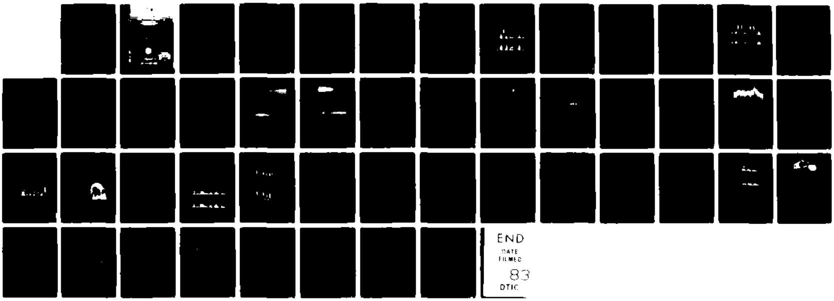
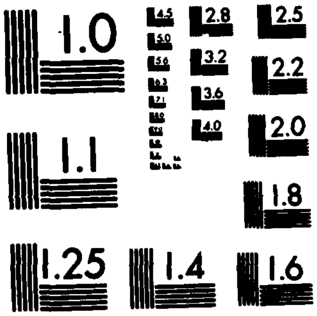
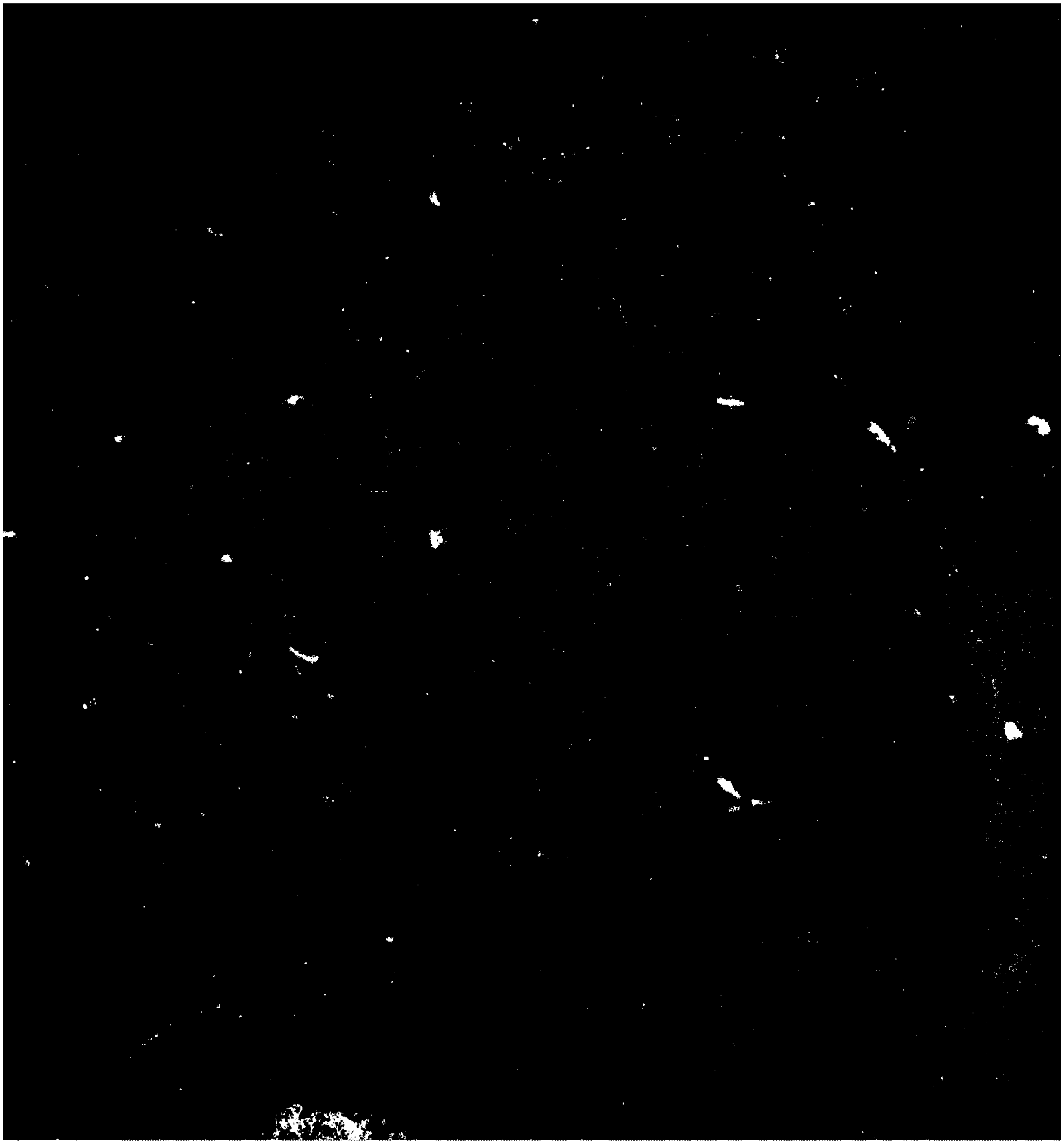
AD-A124 313 IMPROVEMENT OF THE NARROWBAND LINEAR PREDICTIVE CODER 1/1
PART 1 ANALYSIS IMPROVEMENTS(U) NAVAL RESEARCH LAB
WASHINGTON DC G S KANG ET AL. 27 DEC 82 NRL-8645
UNCLASSIFIED SBI-AD-E000 525 F/G 17/2 NL



END
DATE
FILMED
83
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A



SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NRL Report 8645	2. GOVT ACCESSION NO. AD-A124 313	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) IMPROVEMENT OF THE NARROWBAND LINEAR PREDICTIVE CODER, PART 1–ANALYSIS IMPROVEMENTS		5. TYPE OF REPORT & PERIOD COVERED Interim report on a continuing NRL problem 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) G. S. Kang and S. S. Everett		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Research Laboratory Washington, DC 20375		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61153N RR021-05-42 75-1596-0-2
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Washington, DC 22217		12. REPORT DATE December 27, 1982 13. NUMBER OF PAGES 47
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Narrowband speech improvement
Noise-suppression
Automatic gain
Pitch and voicing estimation
Modified linear predictive coding
Regeneration of fricative spectra
Analog front-end

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The narrowband linear predictive coder (LPC) is widely used in both civilian and military applications. Yet, in spite of the many improvements over the years, it is still not universally acceptable to general users. This report presents improvements to five aspects of the LPC analysis which have drawn little attention in the past. These improvements are frequency spreading of sibilant sound spectra, adaptive placement of the analysis window at onsets, modified LPC analysis for sustained vowels, enhancement of LPC spectra in noise, and an automatic gain control. These improvements may be incorporated in the existing narrowband LPC without interfering with the speech sampling rate, the frame rate, or the parameter coding.

CONTENTS

IMPROVEMENT OF THE NARROWBAND LINEAR PREDICTIVE CODER

PART 1—ANALYSIS IMPROVEMENTS

INTRODUCTION

For many years the linear predictive coder (LPC) has been used to convert speech into digital form for secure voice transmission over narrowband channels at low bit rates (less than 5% of the original speech transmission rate). The Navy, as a prime user of narrowband channels for voice communications, has played a significant role in the research and development of LPCs. In 1973 the Navy produced one of the first narrowband LPCs capable of operating in real time. Since 1978 the Navy has been DoD's technical agent for the development of LPCs intended for tri-service tactical use.

Narrowband LPC algorithms have been greatly refined over the years, yet according to recent communicability tests in our branch at the Naval Research Laboratory [1] they are still not as universally acceptable to general users as are commercial telephones. Past attempts in our branch and elsewhere to improve speech quality through minor alterations in one or two areas of the voice-processing algorithms have been largely unsuccessful.

In our present effort to improve the speech quality and acceptability of the narrowband LPC, we have probed some of the more fundamental limitations inherent in the linear predictive analysis and synthesis as applied to narrowband speech coding. Our present effort is part of NRL's continuing efforts to examine the limitations of the narrrowband LPC, to understand peculiarities of human auditory perception, and to find better ways of processing speech for improved quality and intelligibility. We concentrated on improving several specific and interrelated areas that, as our tests show, significantly improve the overall voice quality. The most important part of our effort was to identify and understand critical features that are present in unprocessed speech but absent or suppressed in narrowband LPC speech. We then implemented refinements that compensate for or minimize the discrepancies in these areas.

We have completed our effort on LPC *analysis* improvements, and in this report, Part 1 of a two-part series, we present five areas of LPC analysis improvement. We have effected these improvements without increasing the data rate (2400 bits per second) and without altering the encoding format presently adopted by the Department of Defense [2,3]. The improved narrowband LPC is therefore directly interoperable with other DoD narrowband LPCs under development. Our effort will continue with LPC *synthesis* improvements, the results of which we will document as Part 2 of this two-part series of reports.

OVERVIEW OF OUR LPC ANALYSIS IMPROVEMENTS

We present in the next five sections five topics related to LPC analysis improvement. Each section is self-contained with introductory and background discussion, our technical approach, our test results, and, when appropriate, a summary. We then present a section in which we briefly discuss improvements to other minor areas. The following is an overview of each of these six sections:

Manuscript approved September 8, 1982.

- *Frequency Spreading of Sibilant Sound Spectra.* A large portion of the spectra of voiceless fricatives lies well above the passband of the LPC analyzer. In the absence of these sounds, speech is often difficult to understand. We present a simple method that spreads sibilant sound spectra into the upper end of the passband, providing the listener with sounds similar to those in the original speech.

- *Adaptive Placement of the Analysis Window at Onsets.* Much information relevant to consonant intelligibility is in a small portion of the onset waveform. We present an analysis method that time-aligns the analysis window at this onset.

- *Modified LPC Analysis for Sustained Vowels.* The LPC analysis is based on the assumption that a speech sample can be predicted by a weighted sum of the past samples. However, this principle does not hold well near the beginning of each pitch cycle, where the glottis renews the excitation. We present a method that suppresses this pitch interference.

- *Enhancement of LPC Speech in Noise.* Acoustic-noise interference is a major cause of speech degradation when the LPC is operated in a noisy environment. We examine various factors affecting the LPC performance and describe a preferred noise-suppression method.

- *Automatic Gain Control.* A properly amplified input speech level is essential for the narrowband LPC. In a military environment, though, the narrowband LPC may not always be operated with a well-matched audio input, some audio chains having been designed years ago for analog speech transmission. Manual gain controls in some of the previous voice processors did not work well, because the operators in the field often did not know how to adjust them properly. We describe a software-controlled automatic gain control to compensate for the external gain mismatch.

- *Remarks on the Analog Circuits, the Microphone Shield, and Pitch and Voicing Estimation.* The performance of the narrowband LPC depends highly on the quality of the input speech. We discuss critical aspects of the front-end analog circuits and of the microphone that directly affect the input speech quality. Also, despite marked improvements in pitch and voicing estimation by the current narrowband LPC, occasional errors are still noted. Since powerful processor chips are becoming available, computationally more expensive algorithms should be considered for use in the narrowband LPC while still keeping it small enough to be a part of a voice terminal. We introduce such an algorithm for improved pitch and voicing estimation in the narrowband LPC.

FREQUENCY SPREADING OF SIBILANT SOUNDS

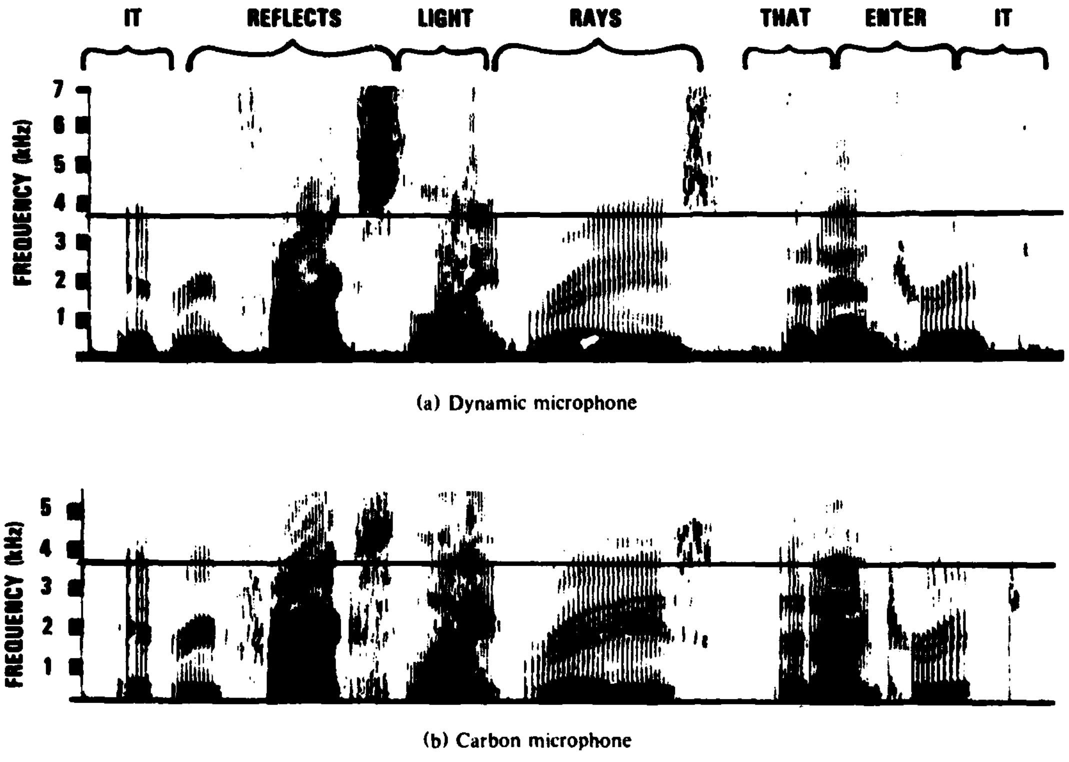
One weakness of the narrowband LPC (and of voice processors in general) is the poor reproduction of voiceless fricative sounds, or sibilants, such as /s/, /sh/, and /ch/. This is because much of the frequency spectra of these sounds lies well above the passband of the voice processor. When these sounds are not reproduced well, it is often difficult to follow what is being said. For example "He's going to sell" may sound like "He's going to hell." Not only is the speech intelligibility degraded, but the speech quality is affected as well, because the absence of the sibilant sounds gives the impression that the talker's teeth are missing. We present a method of spreading the sibilant sound spectra into the passband such that the resulting speech is decidedly more lively and easier to understand.

The Effects of Limited Bandwidth

Since the initial investigation of the vocoder by Dudley in 1939, the front-end bandwidth of the voice processor has typically been 4 kHz. This is a reasonable compromise between the reduction of data rate and the attainment of necessary speech intelligibility. It is adequate for the reproduction of intelligible vowels, but it is not large enough for some voiceless fricative consonants. As illustrated in Fig. 1a, the frequency spectra of these sounds are concentrated above 4 kHz. Once these sounds are eliminated or attenuated at the front end, the loss is irreversable. No algorithmic improvements, including finer coefficient quantization and faster updating of parameters, can restore these sounds. The use of an increased passband would be a simple solution but would undesirably increase the data rate.

Existing Solution for Telephone Speech

The bandwidth of telephone speech is only about 3 kHz, yet the speech quality is widely accepted by general users in normal operational conditions. The solution has been the use of the carbon microphone, initially devised by Edison (who later initially advocated the founding of the Naval Research Laboratory). He devised the carbon microphone in 1877, only 1 year after Bell's invention of the telephone. Bell used a fórm of dynamic microphone. Carbon microphones are still in use in commercial telephones. They produce strong speech signals without additional amplification, but they also disperse the speech spectrum due to the random modulation of the electric resistance caused by the movement of the carbon granules. Figure 1b illustrates the spectrum of the carbon-microphone output. The contrast between Figs. 1a and 1b is particularly evident for sibilant sounds.



(a) Dynamic microphone

(b) Carbon microphone

Fig. 1 — Speech spectra (male voice) from the outputs of dynamic and carbon microphones

Telephone speech, however, is basically an unprocessed analog signal, and the distortions inherent in the carbon-microphone output are not too objectionable to human ears. On the other hand, the narrowband LPC is an analysis/synthesis device, and its performance is highly sensitive to such speech distortions, particularly in vowels. An intelligibility test using the diagnostic rhyme test (DRT) [4] for the 2400-b/s LPC indicates that the overall DRT score is a few points lower with a carbon microphone than with a dynamic microphone [3].

Our Solution

Our method of spreading the sibilant sound spectra exploits the aliasing effect normally suppressed in order to avoid distortion. As a byproduct, it eliminates the need for an elaborately designed antialiasing analog filter, since the necessary filtering is effected by digital computations, which can more readily attain ideal filtering characteristics such as a sharp cutoff rate and a linear phase response over the entire frequency range than can an analog filter. This method is well suited to the digital implementation of voice processors, such as the narrowband LPC.

The conventional front-end processor employs a fixed analog filter that removes the speech contents above 4 kHz. The output of this low-pass filter is then fed into the analog-to-digital converter. Our new front-end processor, on the other hand, uses two low-pass filters (Fig. 2), which are selected adaptively on the basis of the energy distribution in the speech signal. When more energy is below 4 kHz, the normal filter mode is used. The special mode is selected only when a larger proportion of energy is above 4 kHz than is below. At the outset the speech signal is sampled at 16 kHz, double the rate of the conventional approach. A simple 8-kHz analog low-pass filter may be used prior to the analog-to-digital conversion, since little speech energy is present above 8 kHz.

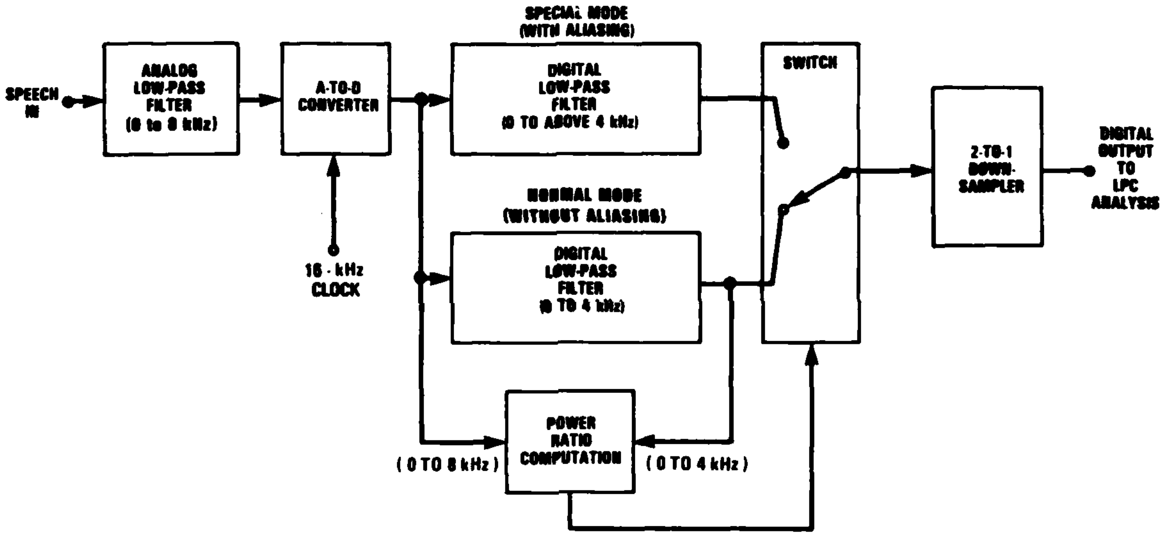

SPECIAL MODE (WITH ALIASING)
SPEECH IN
ANALOG LOW-PASS FILTER (0 to 8 kHz)
A-TO-D CONVERTER
DIGITAL LOW-PASS FILTER (0 TO ABOVE 4 kHz)
SWITCH
2-TO-1 DOWN-SAMPLER
DIGITAL OUTPUT TO LPC ANALYSIS
NORMAL MODE (WITHOUT ALIASING)
16 - kHz CLOCK
DIGITAL LOW-PASS FILTER (0 TO 4 kHz)
POWER RATIO COMPUTATION
(0 TO 8 kHz)
(0 TO 4 kHz)

Fig. 2 — Our front-end processor

Choice of the Two Filters

In our front-end processor both the normal-mode and special-mode low-pass filters are linear phase filters. The impulse response of each filter is expressed by the Hamming-windowed Fourier series

$$h(i) = \begin{cases} G\left[0.54 - 0.46 \cos\left(\frac{2\pi i}{I}\right)\right]\left[0.5 + \sum_{n=1}^{N} \cos\left(\frac{n\pi i}{I} - 0.5\right)\right], & \text{for } 0 \leqslant i \leqslant I - 1, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where the factor G makes the sum of the impulse response samples unity (a dc gain of unity). The quantity I is the total number of impulse response samples, which is related to the attenuation rate beyond the cutoff frequency. The quantity N is related to the cutoff frequency for a given value of I. The impulse response is symmetric with respect to the midpoint. Thus the phase response is linear.

The characteristics of the normal-mode low-pass filter are comparable to those of the antialiasing low-pass filter now used in the narrowband LPC. Thus, the −3-dB cutoff frequency is at approximately 4 kHz, and the cutoff rate is of the order of −100 dB/octave. Such a filter may be realized by letting $I = 43$ and $N = 22$ in Eq. (1). The frequency response of this filter above its cutoff frequency is listed in Table 1. The magnitude of the maximum in-band ripple is 0.01 dB.

We chose filter B (Table 1) as the special-mode low-pass filter because, after repeated listening tests with the sentences in Table 2, we decided that the most natural voiceless fricative sounds were produced with this filter for both male and female voices. This filter is obtained by letting $I = 11$ and $N = 7$ in Eq. (1). The maximum in-band ripple of this filter is 0.04 dB. The impulse responses of the normal-mode low-pass filter and the selected special-mode low-pass filter are listed in Table 3.

Table 1 — Cutoff Characteristics of the Four Low-Pass Filters Tested as the Choice for the Special-Mode Filter

Freq. (Hz)	Frequency Response of the Normal-Mode Filter (dB)	Response of Special-Mode Filter A, B, C, or D (dB)			
		A	B*	C	D
4000	−3.97	−1.39	−1.41	−3.91	−4.03
5000	−56.25	−5.00	−6.00	−10.26	−12.25
6000	−62.87	−12.23	−16.80	−22.97	−30.52
7000	−72.46	−26.24	−47.85	−48.44	−56.05

*Preferred choice based on listening tests with the sentences in Table 2.

Table 2 — Sentences

Her purse was full of useless trash.
Be sure to see the side show.
She sews nice and straight.
Once we stood beside the shore.
Sue washed the spoons and forks.

Table 3 — Impulse Responses of the Two Digital Low-Pass Filters Shown in Fig. 1

Index j	Response $h_1(j)$	Index j	Response $h_1(j)$	Index j	Response $h_1(j)$	Index j	Response $h_1(j)$	Index j	Response $h_2(j)$
Normal-Mode Filter								Special-Mode Filter	
1 and 43	0.00103	7 and 37	−0.00557	13 and 31	0.02282	19 and 25	−0.10113	1 and 11	−0.00370
2 and 42	0.00112	8 and 36	−0.00396	14 and 30	0.00829	20 and 24	−0.01113	2 and 10	0.02577
3 and 41	−0.00171	9 and 35	0.00930	15 and 29	−0.03505	21 and 23	0.31613	3 and 9	−0.02245
4 and 40	−0.00174	10 and 34	0.00538	16 and 28	−0.00954	22	0.51046	4 and 8	−0.08744
5 and 39	0.00314	11 and 33	−0.01479	12 and 27	0.05581			5 and 7	0.027607
6 and 38	0.00271	12 and 32	−0.00687	18 and 26	0.01052			6	0.62348

Front-End Processing

The front-end processor performs the normal mode of low-pass filter operation. Thus, the filter output is expressed by

$$y(i) = \sum_{j=1}^{43} x(i-j)h_1(j)$$
$$= x(i-22)h_1(22) + \sum_{j=1}^{21} [x(i-j) + x(i-44+j)]h_1(j), \tag{2}$$

where $x(j)$ is the input and $y(i)$ is the normal-mode low-pass filter output. The term $h_1(j)$ is the impulse response of the normal-mode low-pass filter, as listed in Table 3. At the same time, both the input and output energies are computed by

$$P_x(i) = P_x(i-1) + [x^2(i) - P_x(i-1)]/32 \tag{3a}$$

and

$$P_y(i) = P_y(i-1) + [y^2(i) - P_y(i-1)]/32. \tag{3b}$$

The single-pole filter employed in Eqs. (3) has a −3-dB cutoff frequency at 40 Hz which has proven to be satisfactory for the present application.

If the total speech power $p_x(i)$ is less than twice the partial speech power $p_y(i)$, the output of the normal-mode filter is down-sampled by a factor of 2 and passed on to the LPC analysis. On the other hand, if $p_x(i)$ is greater than $p_y(i)$, the special mode is chosen, and the following output is down-sampled by a factor of 2 and then passed on to the LPC analysis:

$$y(i) = x(i-22)h_2(6) + \sum_{j=1}^{5} [x(i-16-j) + x(i-28+j)]h_2(j), \tag{4}$$

where $h_2(j)$ is the impulse response of the special-mode low-pass filter as listed in Table 3. To make a smooth transition from one mode to another, both filter outputs are time-aligned at 22 sampling time intervals behind the present time (the midpoint of the normal-mode filter's impulse response).

Justification of Our Approach

The way the detection threshold is set, the only sounds which trigger the special filter are /s/ and sometimes for high-pitched voices /sh/ and /ch/. The spectra of the weaker fricatives /f/ and /th/ are more spread out and do not have sufficient energy to be detected in this way. Voiced fricatives and affricates, including /z/ and /zh/, are not aliased for two reasons: they have a relatively strong low-frequency voicing component, and they have less energy above 4 kHz than do /s/ and /sh/.

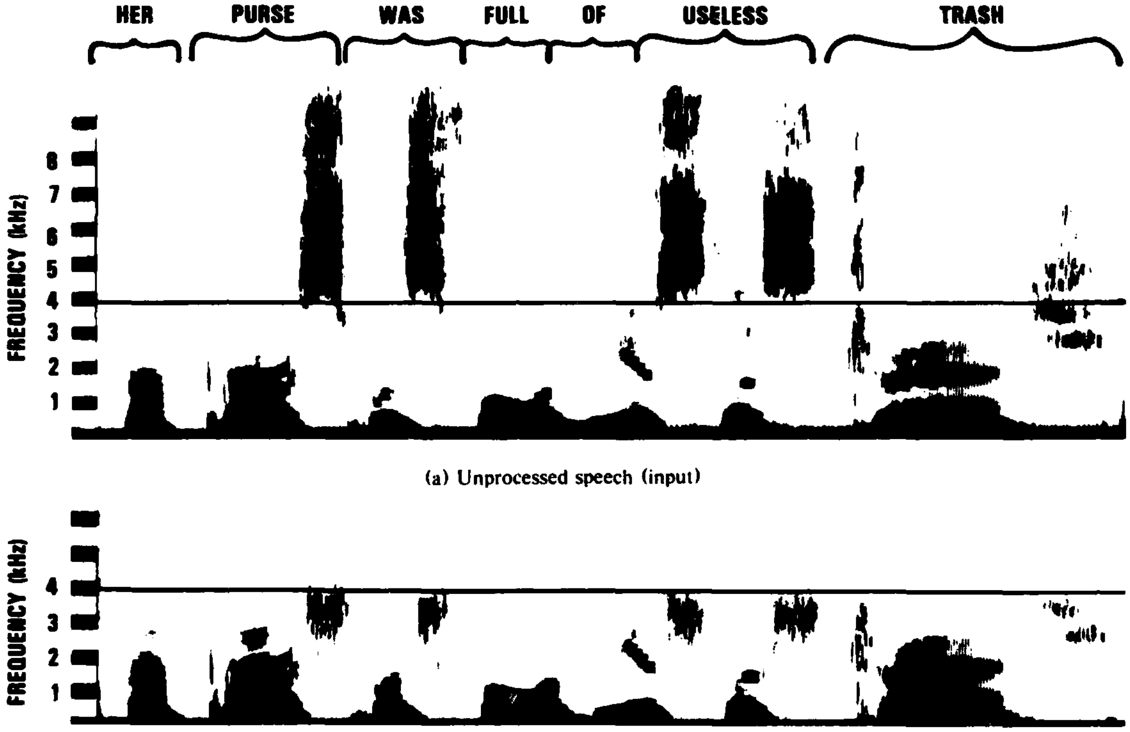
Obviously the changes in the frequency spectra of the aliased fricatives will cause them to sound somewhat different. Studies of fricative spectra have shown that /s/ usually has little energy below 4 kHz, particularly with high-pitched voices [5,6]. The fricative /sh/, on the other hand, typically has a strong component between 2 and 3 kHz for males and only slightly higher for females. Therefore we chose the special-mode filter such that the aliasing process does not reflect the spectra of /s/ much below 3 kHz.

Studies using synthetic fricatives with ambiguous spectra have shown that the vowel quality and formant transitions present in the neighboring vocalic segments have a strong influence on the identification of the frication [7,8]. In general the ambiguous fricatives are heard as /s/ in the context of /s/ transitions and as /sh/ in the context of /sh/ transitions. The aliased /s/ spectra are in essence ambiguous fricatives, since they lack the distinguishing characteristics of either /s/ or /sh/. The aliasing process does not affect other sounds, however, and the vowel quality and formant transitions are, for the most part, preserved in the LPC processing.

On the whole, then, the aliasing of the fricative spectra should not interfere with the discrimination of these sounds. This is particularly true in a conversational environment, where much contextual information is available to the listener to help in discriminating the sounds.

Summary

Figure 3 shows the input and output spectra of our front-end processor. As apparent from Fig. 3b, the spectra of sibilant sounds are dispersed into the passband, mainly above 3 kHz. For other sounds there is virtually no distortion, unlike with the carbon microphone (Fig. 1).



(a) Unprocessed speech (input)

(b) Speech with selective aliasing (output)

Fig. 3 — Input and output spectra (female voice) of our front-end processor

The filter-selection logic works well despite its simplicity, the logic being to select the special-mode low-pass filter only when the total speech power (0 to 8 kHz) is at least twice the partial speech power (0 to 4 kHz). Thus aliasing takes place only for those sounds having predominantly high-frequency components. Ambient acoustic noise could possibly interfere with the filter-selection logic, but the noise would have to have a strong component between 4 and 5 kHz and little energy below 4 kHz. None of the military noise platforms that we have investigated (including tanks, helicopters, ships, and high-performance jets) exhibit these characteristics.

As shown in Figs. 1 and 3, the fricative spectra of both male and female voices are concentrated above the LPC passband. Thus aliasing of the fricative spectra is expected to improve the intelligibility of narrowband speech in general. In support of this expectation a diagnostic rhyme test (DRT) for one female speaker (Table 4) shows an overall gain of 5 points (from 79.2 to 84.0) when the fricative spectra are aliased.

Table 4 — DRT Scores for One Female Speaker Realized Through the Use of Aliasing

Sound Class	Score		
	Without Aliasing	With Aliasing	Change
Voicing	74.2	88.0	+13.8
Nasality	98.4	96.1	−2.3
Sustention	67.2	77.1	+9.9
Sibilation	78.1	82.6	+4.5
Graveness	68.7	71.1	+2.4
Compactness	88.3	88.8	+0.5
Overall	79.2	84.0	+4.8

ADAPTIVE PLACEMENT OF THE ANALYSIS WINDOW AT ONSETS

It is well known that the narrowband LPC, like narrowband voice processors in general, does not reproduce speech onsets clearly—they sound slurred and fuzzy. This effect becomes more pronounced when the speech is contaminated by background noise. Even with an almost inaudible amount of noise mixed with the speech, the narrowband LPC output often sounds like a drunken voice. A major factor contributing to this poor speech quality is the improper placement of the analysis window at onsets, where the speech characteristics change drastically in a short time. When the critical information contained in the onset is missed, the resulting speech will be slurred and difficult to understand. We present an analysis method that provides information as to where the analysis window should be placed to minimize this distortion.

Deficiencies in the Current LPC Analysis

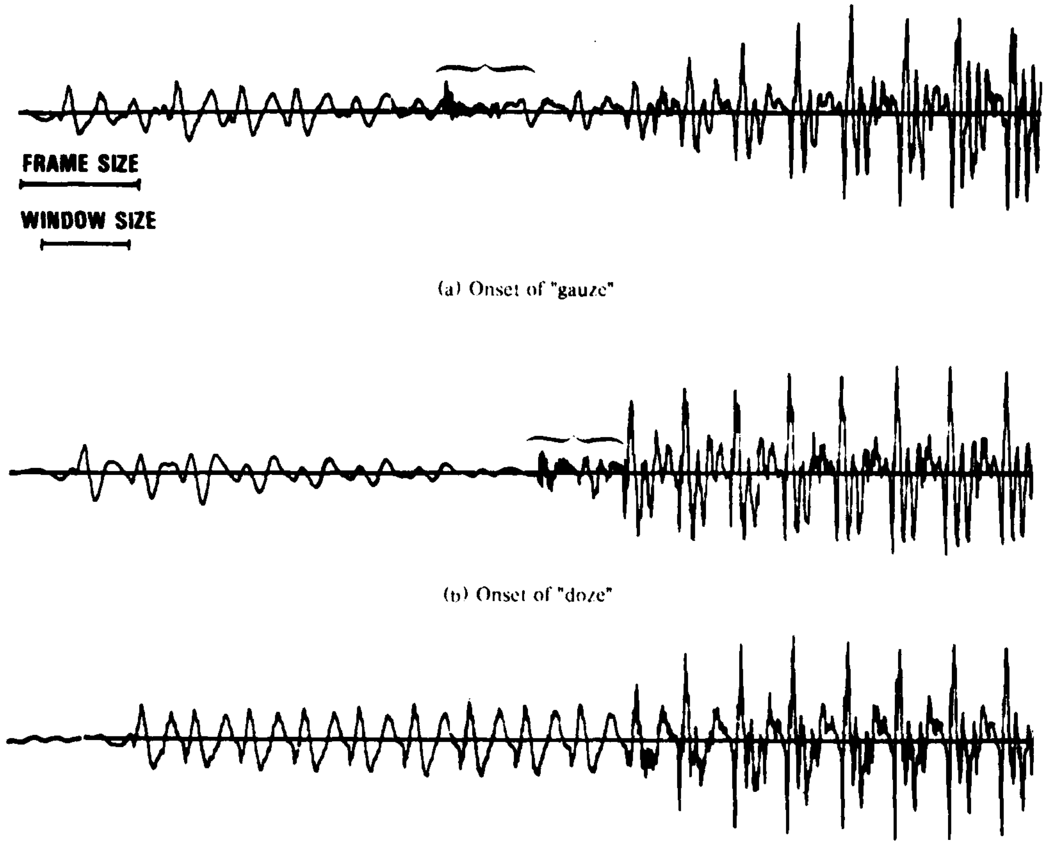
All narrowband voice processors, including the LPC, exhibit marked deficiencies in the reproduction of some abrupt voice onsets, including /b/, /d/, and /g/. Much research has been devoted to the study of these sounds and their unvoiced cognates /p/, /t/, and /k/ [9,10]. It has been shown that the initial 10 to 20 ms of a stop consonant provide the principal cues to place of articulation [11]. If the information contained in this critical interval were distorted, it should lead to poor identification and discrimination for these sounds. From the early days of narrowband LPC development, the diagnostic rhyme test (DRT) has consistently indicated that this is indeed the case.

Unlike wideband voice processors, which transmit the speech waveform (or waveforms derived from speech), the narrowband LPC transmits encoded speech parameters. These parameters are extracted once per frame from the speech samples within the analysis window. The frame size used is the current DoD standard of 22.5 ms, but the window has been made smaller—16.25 ms—to reduce the smudging effect. The window epoch is flexible, and it may be placed anywhere within or slightly outside the current frame.

None of the current narrowband LPCs perform any analysis prior to placing the analysis window. Although some of the onset distortion is due to the inability of the limited number of poles in the basic LPC speech model to adequately describe complicated onset waveforms, much of the distortion is due to this random placement of the analysis window. Since the analysis window is smaller than the frame size and can be placed anywhere within or slightly outside the current frame, the chance is good that the window will be positioned improperly without an appropriate analysis beforehand.

Effects of Analysis-Window Misplacement at Onsets

One form of window misplacement results in the omission of the first few milliseconds of the onset waveform that are so critical to consonant intelligibility. This critical section of the speech waveform is often little more than a short burst of noise-like signal immediately preceding the start of the vowel, as indicated by braces in Fig. 4, and is often shorter than the window. Thus, even if the analysis window is properly placed, the averaging process in the LPC analysis tends to diminish the effect of the burst waveform. Once this critical information is lost, both "gauze" and "doze" tend to sound like "nose."



(a) Onset of "gauze"

(b) Onset of "doze"

(c) Onset of "nose"

Fig. 4 — Waveforms of speech onsets

Another form of window misplacement leads to interference between two adjacent sound elements. Consider the onset waveform of "think," where the unvoiced sound /th/ is followed by the voiced sound /i/, as shown in Fig. 5a. In this case the proper window positions are at location 1, over the unvoiced sound, and at location 3, over the voiced sound. The resulting synthesis-filter responses (which responses are used in the speech synthesizer, to be described in Part 2 of this two-part series of reports) are shown in Figs. 5b and 5d, respectively. The analysis window at location 2 produces an interference, because it includes portions of both the unvoiced and the voiced sounds. The resulting filter response, shown in Fig. 5c, is inappropriate because it represents neither /th/ nor /i/. It will produce a distorted sound that lasts for one full frame. Stated another way, it will take an additional frame for the synthesis filter to produce the proper sound. Consequently, the processed speech sounds sluggish, and the lively quality of the original speech is greatly diminished. This type of interference could occur two or three times within a single word (such as "station," "repetition," "statistics," and "exercise").

Interference can also occur when the stop consonant is preceded by a vowel. The adverse effect on the filter coefficients is even more severe in this case because of the strong resonant frequency components of the vowel. One example is the word "body," where the consonant /d/ is preceded by the vowel /o/. If the analysis window contains both /d/ and /o/, "body," may be heard as "boy" at the narrowband LPC output. Spectrographic analysis indicates that the formant discontinuities at the onset of /d/ are definitely not as distinct in the processed waveform as in the unprocessed speech. Interference of this kind would be expected to occur frequently in normal conversation, since no one puts a distinctive gap prior to each consonant.

Proper windowing is even more essential with noisy speech. Satisfactory operation of the narrowband LPC in a platform where the noise level is greater than 80 dB requires some sort of noise-suppression processing, such as we present later in this report, prior to the LPC analysis. Without proper windowing, however, noise suppression alone has shown little promise.

Our Onset Detector Algorithm

We now describe our analysis procedure for detecting the critical speech transitions where the analysis window should be placed. The method performs a single-order prediction in running time for the preemphasized speech waveform. We use the difference between the forward and backward prediction coefficients, and their time histories, for the onset detection.

The forward prediction coefficient minimizes the forward prediction error in the mean-square sense. Likewise, the backward prediction coefficient minimizes the backward prediction error in the mean-square sense. If the speech is stationary, both errors have the identical mean-square value, and the two coefficients are equal. Since speech statistics change rapidly at the onset (the speech is not stationary), the forward prediction coefficient will differ from the backward prediction coefficient. We use the difference, observable at each sampling time instant, for the onset detection.

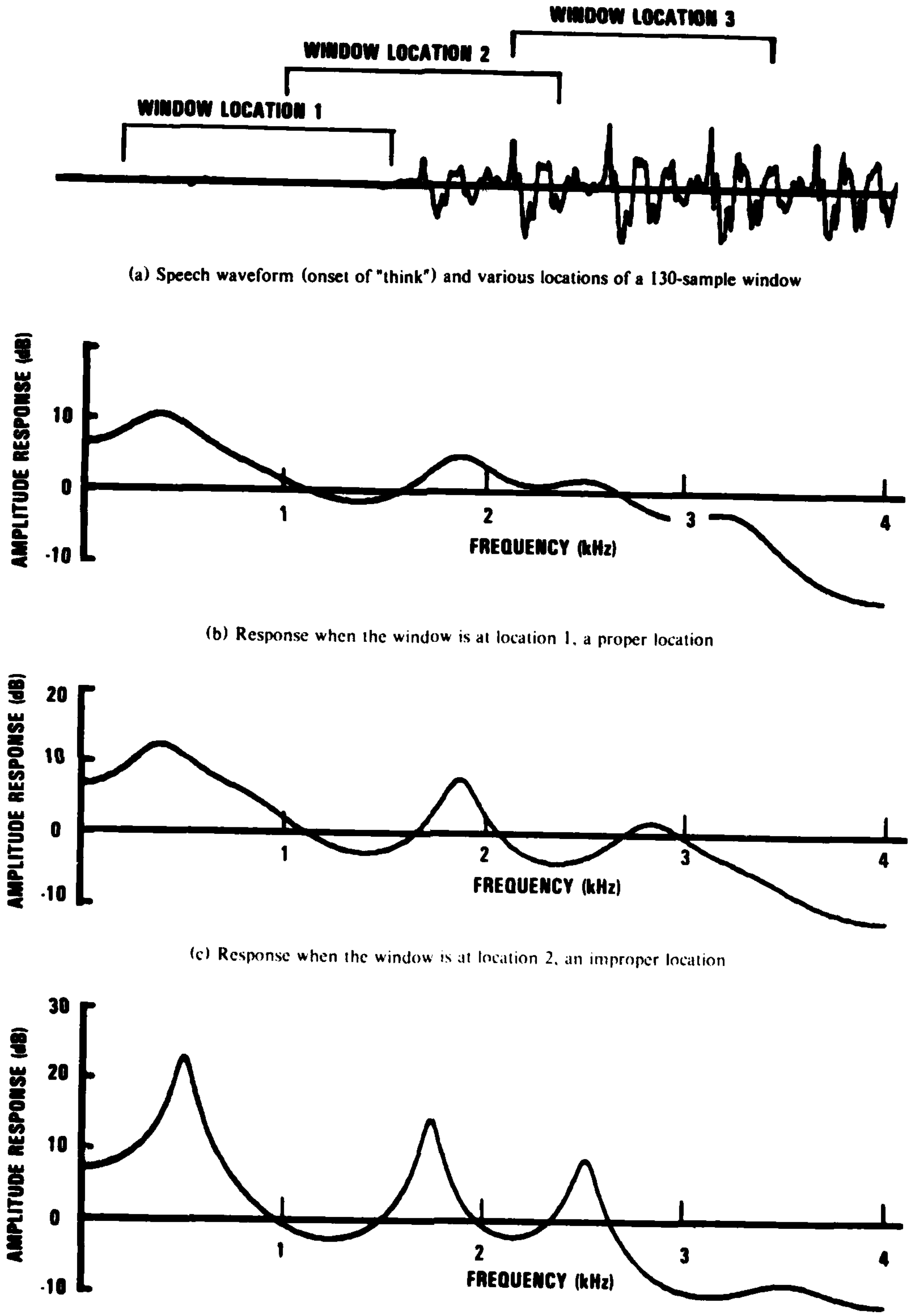
In addition we use the time histories of both coefficients. If the speech has predominately high-frequency components, the coefficients will be close to -1. On the other hand, if the speech contains predominately low frequency components, the coefficients will be close to $+1$. Thus the coefficient indicates the general nature of the speech spectral tilt. Since the tilt changes drastically at the onset, particularly when going from an unvoiced state to a voiced state, the change in both prediction coefficients in running time is an excellent means of onset detection.

The detection algorithm is computationally efficient, requiring three multiplications, three summations, and two divisions for each speech sample. The preemphasis operation is not included here, since it is required for the LPC with or without the onset detector. As usual, windows are placed semi-pitch-synchronously if the speech is voiced and at fixed time intervals if the speech is unvoiced. The time marker generated by the onset detector is used to rephase the windowing cycle.

The onset detector exploits the flow-form LPC analysis previously implemented by the Navy [12]. This approach provides a sample-by-sample time history of prediction coefficients. Initially each speech sample is preemphasized in running time by a factor identical to that required by the DoD narrowband LPC. Thus

$$x_i = s_i - (15/16)s_{i-1}, \qquad (5)$$

where s_i is a given speech sample and x_i is the preemphasized speech sample.



(a) Speech waveform (onset of "think") and various locations of a 130-sample window

(b) Response when the window is at location 1, a proper location

(c) Response when the window is at location 2, an improper location

(d) Response when the window is at location 3, a proper location

Fig. 5 — Amplitude responses of the synthesis filter for various analysis-window locations near a speech onset

The forward prediction error $\epsilon_{f,i}$ from the first-order predictor is

$$\epsilon_{f,i} = x_i - k_{f,i} x_{i-1}, \tag{6}$$

where $k_{f,i}$, which is the forward prediction coefficient and which minimizes $\epsilon_{f,i}$ in the mean-square sense, is

$$k_{f,i} = \frac{\overline{x_i x_{i-1}}}{\overline{x_{i-1}^2}}. \tag{7}$$

The bar signifies low-pass filtering by a single-pole filter that has a feedback constant of 63/64. The −3-dB cutoff frequency of this filter is 20 Hz.

Likewise, the backward prediction error $\epsilon_{b,i}$ from the first-order predictor is

$$\epsilon_{b,i} = x_{i-1} - k_{b,i} x_i, \tag{8}$$

where $k_{b,i}$, which is the backward prediction coefficient and which minimizes $\epsilon_{b,i}$ in the mean-square sense, is

$$k_{b,i} = \frac{\overline{x_i x_{i-1}}}{\overline{x_i^2}}. \tag{9}$$

At every 16th sampling time interval a check is made for the following two conditions:

- The sample-by-sample difference between the forward and backward prediction coefficients has exceeded the prescribed threshold (L1) one or more times in the past 16 sampling time intervals;

- The total change in either the forward or backward prediction coefficient for the past 16 sampling time-intervals exceeds the prescribed threshold level (L2).

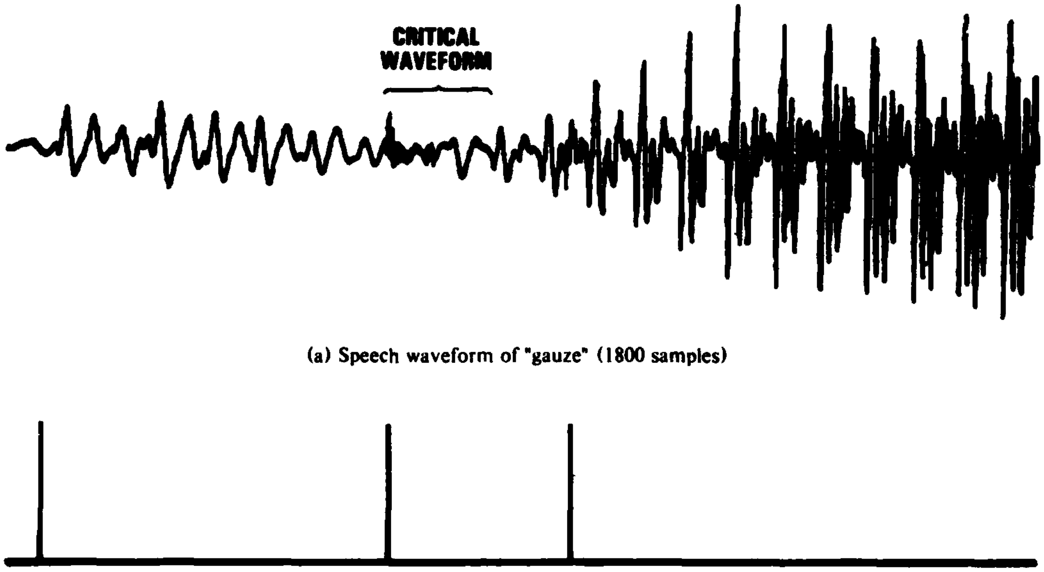
If one or both of these conditions is met, the detector generates a time marker. According to experimentation with a variety of waveforms, a threshold level of 0.25 is acceptable for both L1 and L2.

The leading edge of the analysis window should be placed coincident with or slightly forward of the time marker to compensate for the rise time of the low-pass filter. If the onset detector registers significant speech transitions at two consecutive observation intervals, indicating that the speech has changed significantly over 32 sampling time intervals, the analysis window is aligned with the first time marker.

Test Results

Our analysis is the first since the initial development of the narrowband LPC a decade ago to place windows properly at critical speech transitions. Figures 6 through 9 show four speech waveforms and the output of the onset detector, as plotted by the computer. Each example has a special characteristic. The examples are (Fig. 6) the speech waveform of "gauze" with a short burst of critical waveform, (Fig. 7) the speech waveform of "think" with an unvoiced sound followed by a voiced sound, (Fig. 8) the speech waveform of "bodies" with a consonant preceded by the trailing end of a vowel, and (Fig. 9) the speech waveform of "Bob" with interference by helicopter noise, at a sound pressure level (SPL) of 115 dB. The results are highly promising. The onset detector performed as expected for these four examples and for other speech samples tested.

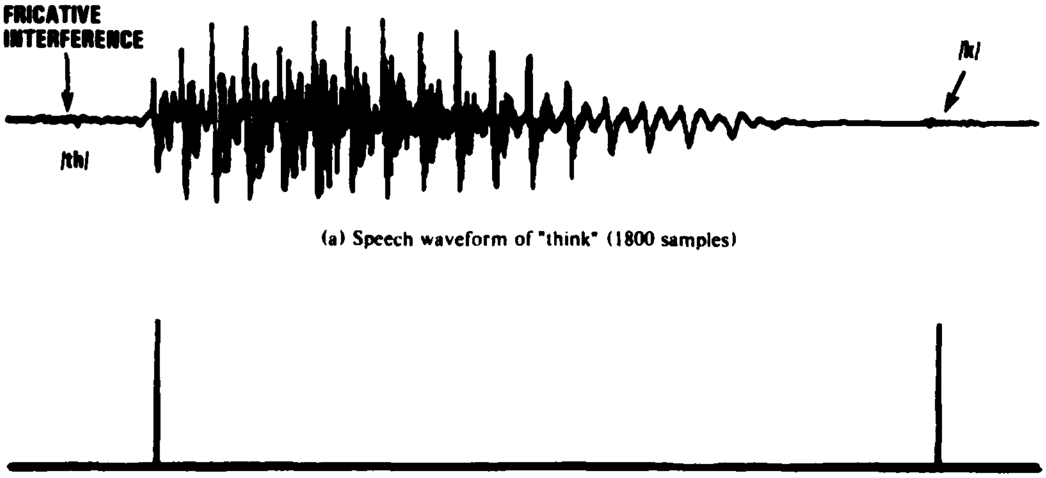
To further validate these findings, we made a one-male DRT source tape in a tank platform and processed it in real time through the existing 2400-b/s LPC, both with the adaptive analysis window placement and without. The DRT is an ideal means of evaluating the onset reproduction, because it



(a) Speech waveform of "gauze" (1800 samples)

(b) Output of the onset detector

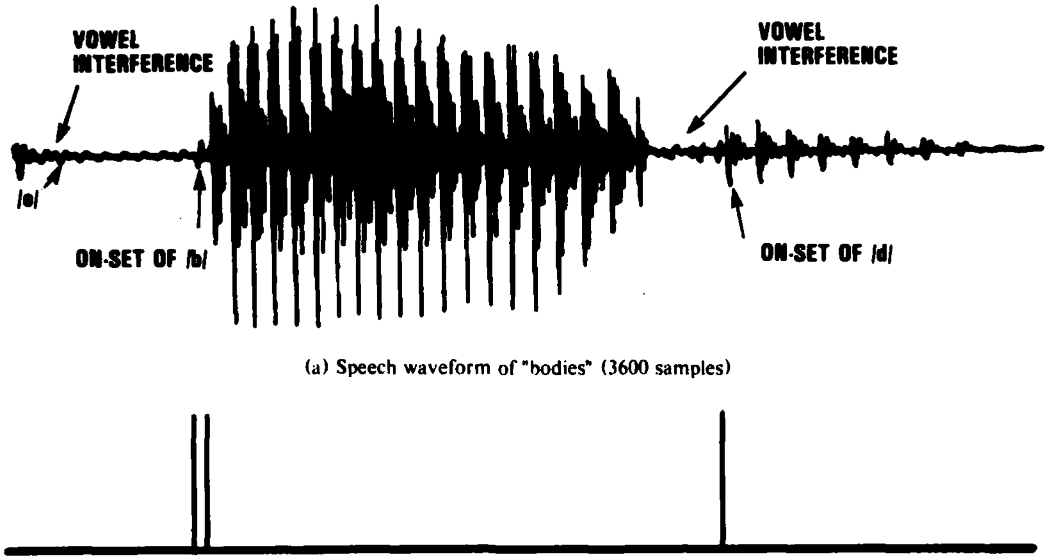
Fig. 6 — Performance of the onset detector in the presence of a critical onset waveform



(a) Speech waveform of "think" (1800 samples)

(b) Output of the onset detector

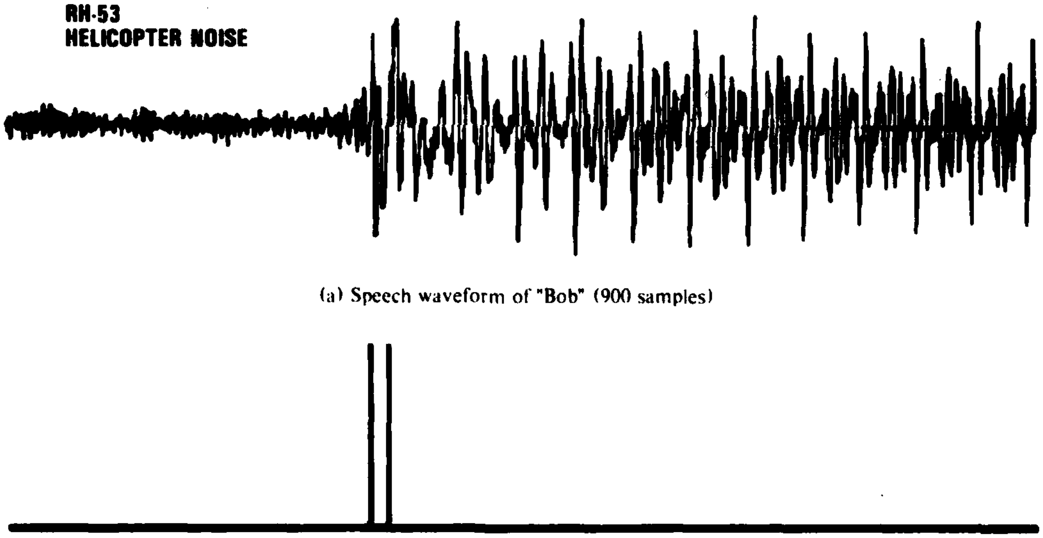
Fig. 7 — Performance of the onset detector in the presence of fricative interference



(a) Speech waveform of "bodies" (3600 samples)

(b) Output of the onset detector

Fig. 8 — Performance of the onset detector in the presence of vowel interference



(a) Speech waveform of "Bob" (900 samples)

(b) Output of the onset detector

Fig. 9 — Performance of the onset detector in the presence of severe background noise

tests initial consonants. It is encouraging that the adaptive window placement produced a 14-point improvement for the attribute "nasality" (which tests the contrasts /n/ vs /d/, and /m/ vs /b/). Likewise, a 10-point improvement was shown for the attribute "graveness" (which tests /b/ vs /d/, /p/ vs /t/, and like contrasts), and an 8-point improvement was evident for the attribute "compactness" (which tests /g/ vs /d/, /k/ vs /t/, and like contrasts). The overall score of 72.0 was raised to 75.4 with the implementation of the adaptive window placement at onsets.

MODIFIED LPC ANALYSIS FOR SUSTAINED VOWELS

The sustained vowel waveform is quasiperiodic at the pitch frequency. Therefore the synthesis filter response, as determined by LPC coefficients, varies quasiperiodically at the pitch frequency. Thus, unless the analysis window is placed synchronously with the pitch cycle, the synthesized speech becomes warbly due to filter-response variations from frame to frame. This interframe pitch interference has attracted much attention in the past [13], and the solution has been to place an analysis window at an integer multiple of the pitch period from the preceding window location. This method, used in the DoD narrowband LPC, has proven to be effective for minimizing warble in sustained vowels.

Our investigation, however, has been concerned with another form of pitch interference—intraframe pitch interference—which has been neglected in the past. The LPC analysis is based on the assumption that a given speech sample can be predicted by a weighted sum of past samples, and a set of weighting factors (or LPC coefficients) is estimated by way of minimizing the mean-square prediction errors. This principle does not hold well near the pitch epoch, though, where the ongoing waveform is disturbed by the glottis excitation. The current LPC analyzer is a simple unbiased estimator that cannot discard those prediction equations contributing to much larger prediction errors than others. Inclusion of these outliers in the LPC analysis leads to broadened resonant bandwidths that make the synthesized speech fuzzy. In the following subsection we present a way of reducing this intraframe pitch interference.

Our Modified LPC Analysis

Our modified LPC analysis is in essence a two-path analysis. The first-path LPC analysis is identical to the conventional LPC analysis: prediction coefficients are derived by solving a set of overdetermined, simultaneous equations (typically 120 equations with ten unknowns). In the second-path LPC analysis some of the prediction equations contributing to large prediction errors are eliminated. As we will show, the eliminated equations are those which predict speech samples near the pitch epoch.

The first-path LPC analysis begins by preemphasizing the speech with a single-zero filter having a zero at $z = 15/16$. (Hereafter speech samples, residual samples and related illustrations are results of this preemphasis.) As usual, a speech sample x_i is represented by a weighted sum of past samples:

$$x_i = \sum_{n=1}^{N} \alpha_n x_{i-n} + \epsilon_i, \qquad i = N+1,\ N+2,\ \ldots,\ I, \tag{10}$$

where α_n is the nth prediction coefficient and ϵ_i is the ith prediction residual. For the DoD narrowband LPC, $N = 10$ and $I = 130$, and the number of simultaneous equations is 120. The index i is consecutive for the conventional LPC analysis. In matrix notation, Eq. (10) may be written as

$$\mathrm{X} = \mathrm{HA} + \mathrm{E}, \tag{11}$$

and the solution for prediction coefficients by the least-squares method [14] is

$$\hat{\mathrm{A}} = (\mathrm{H}^{\mathrm{T}}\mathrm{H})^{-1}(\mathrm{H}^{\mathrm{T}}\mathrm{X}). \tag{12}$$

Since H contains consecutive speech samples in each row or column, computations in $\mathrm{H}^{\mathrm{T}}\mathrm{H}$ can be drastically simplified, since each element of $\mathrm{H}^{\mathrm{T}}\mathrm{H}$ can be obtained iteratively [3]. The inversion $\mathrm{H}^{\mathrm{T}}\mathrm{H}$

may be effected more conveniently through Cholesky decomposition [2,3,15]. The conventional LPC analysis ends with Eq. (12).

In our modified LPC analysis the second-path analysis begins with computations for prediction residual (or error) samples based on the prediction coefficients obtained by the first-path analysis. Since in general a speech sample cannot be perfectly represented by a weighted sum of past samples, the error is not expected to be zero. The error, however, becomes particularly large near the pitch epoch, where the glottal excitation is renewed. If the error is larger than the prescribed threshold level, the corresponding prediction equation will be excluded from the second-path LPC analysis.

According to experimentation, the error threshold level can be anywhere between 1.5 and 2.5 times the root-mean-square (RMS) value of the residual samples within the frame. If the threshold level is chosen to be twice the RMS value, on the average 95% of the residual magnitudes will fall under the threshold level. This figure is obtained from the probability density function of residual samples (Fig. 10) of 15,000 voiced frames from both male and female voices, where each frame contains 120 residual samples.

The elimination of a prediction equation does not mean the elimination of a speech sample. Unless prediction errors exceed the threshold level at $N + 1$ consecutive sample time intervals—which is unlikely—none of the speech samples will be eliminated.

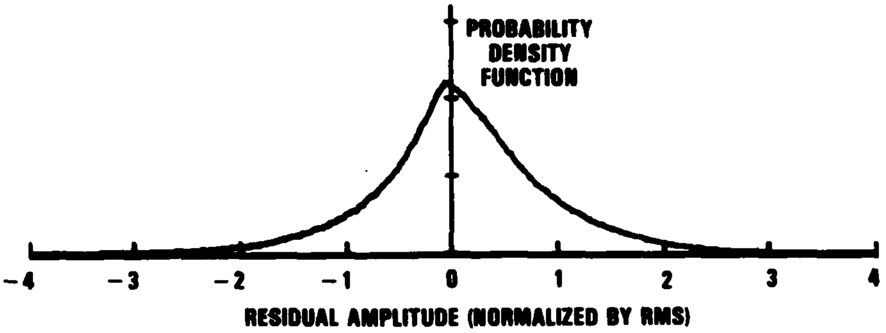
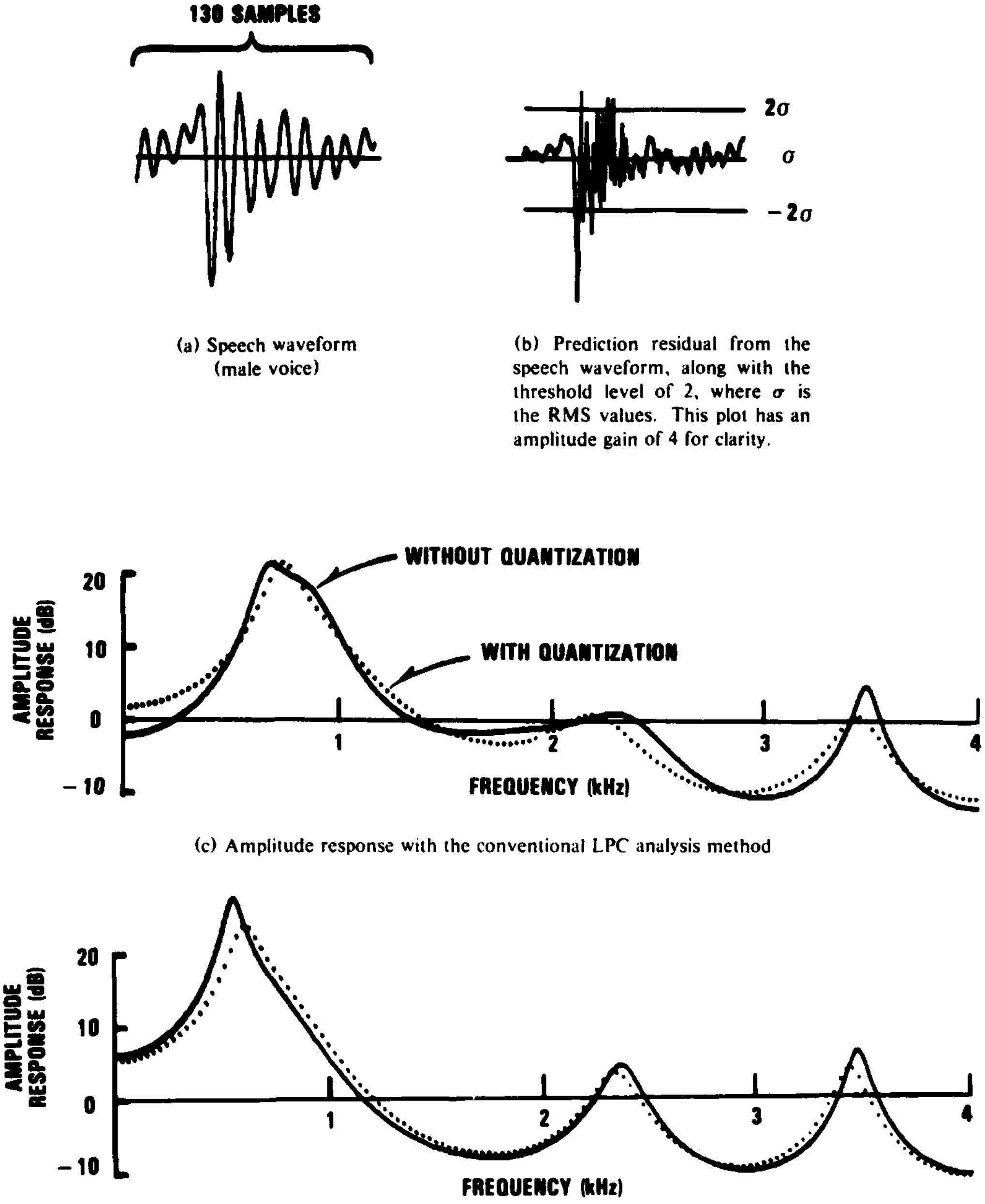


Fig. 10 — Probability density function of prediction residuals (from voiced speech samples)

The set of prediction equations for the second-path LPC analysis is given by Eq. (10), except that the index i is in general not consecutive. The solution for the prediction coefficients is again given by Eq. (12). The computations in H^TH cannot be simplified as in the first-path LPC analysis, because the index i is not consecutive. Thus it is obtained by multiplying H^T by H as indicated. The computational procedure for inverting H^TH for the second-path LPC analysis, however, is identical to that of the first-path LPC analysis.

Examples

Figure 11 illustrates the effectiveness of the modified LPC analysis. Figure 11a shows 130 unprocessed speech samples selected for the LPC analysis. As is often the case with male voices, there is only one pitch epoch in the frame. Figure 11b shows the prediction residual samples, amplified 4 times for display purposes. The threshold level equal to twice the RMS value is also indicated. Figure 11c is the speech-synthesis-filter amplitude response computed from the LPC coefficients generated by the conventional LPC analysis. The bandwidths of both the first and second of the three formants are unusually broad, which makes the synthesized speech sound fuzzy. The solid line indicates the synthesis-filter amplitude response with unquantized coefficients, and the dotted line indicates the response with the coefficients quantized in the manner specified by the DoD narrowband LPC [2,3].



(a) Speech waveform (male voice)

(b) Prediction residual from the speech waveform, along with the threshold level of 2, where σ is the RMS values. This plot has an amplitude gain of 4 for clarity.

(c) Amplitude response with the conventional LPC analysis method

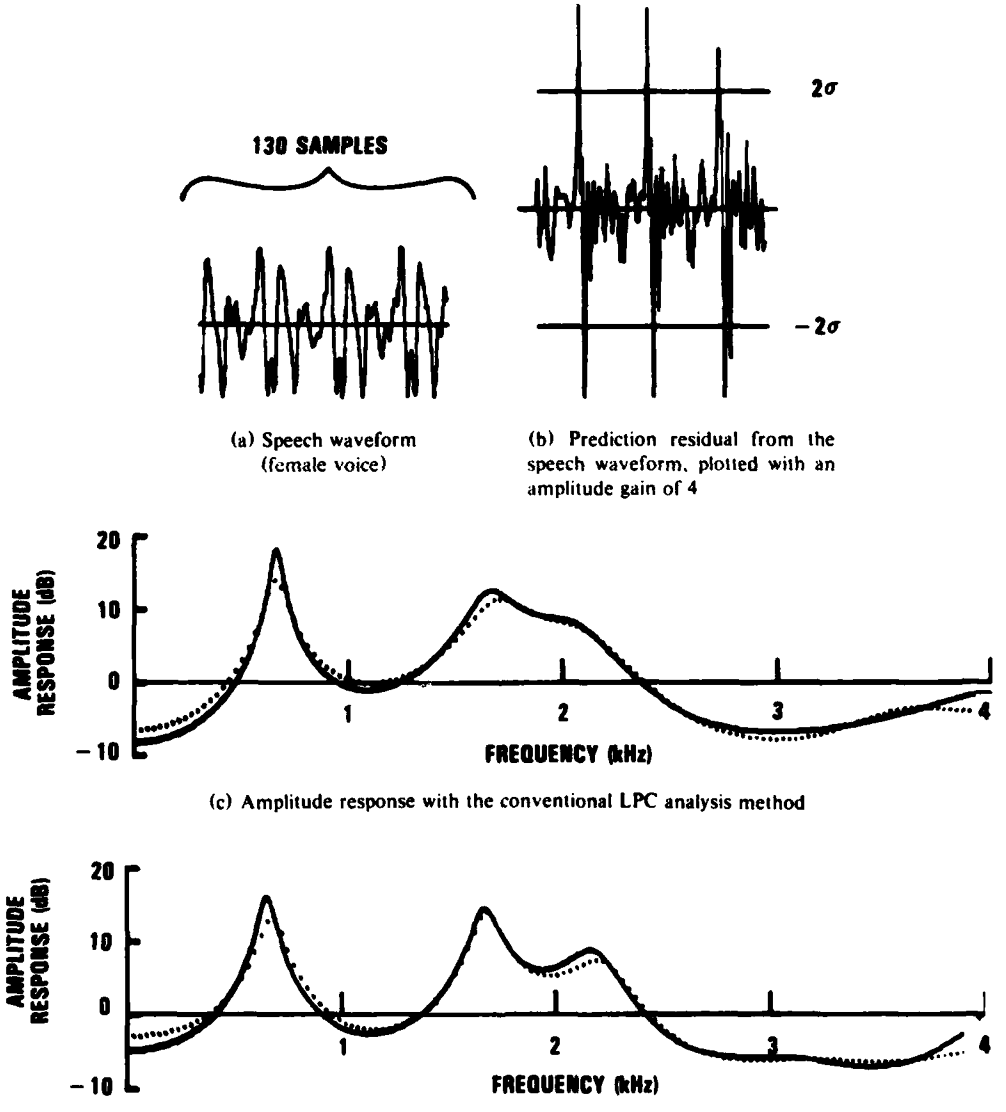
(d) Amplitude response with our modified LPC analysis method

Fig. 11 — Synthesis-filter amplitude responses with the conventional LPC analysis method and with our modified method, with a male-voice input

Figure 12 is another example, using a female voice. Here three pitch epochs are in the frame. Our modified LPC analysis resolves the frequency of the third formant much more sharply than does the conventional LPC analysis.

Summary

Our modified LPC analysis definitely focuses speech sounds more than the conventional LPC analysis does. The one drawback is the increased computational load. The modified approach requires not only the matrix loading and inversion required by the conventional LPC analysis but an additional matrix inversion identical to that of the conventional approach, further residual computations, and far more involved matrix loading.



(a) Speech waveform (female voice)

(b) Prediction residual from the speech waveform, plotted with an amplitude gain of 4

(c) Amplitude response with the conventional LPC analysis method

(d) Amplitude response with our modified LPC analysis method

Fig. 12 — Synthesis-filter amplitude responses with the two LPC analysis methods, with a female-voice input

Conceivably the first-path LPC analysis could be bypassed by preselecting speech samples based on a knowledge of the locations of the pitch epochs. Further investigation of this approach will be conducted when the modified LPC analysis is implemented for real-time operation. The preliminary results, however, show a definite promise for improving the speech quality of the narrowband LPC.

ENHANCEMENT OF LPC SPEECH IN NOISE

Interference by acoustic noise is a major cause of speech degradation in all digital voice processors. In an extremely noisy platform, such as a tank, the intelligibility level of narrowband LPC speech is often unacceptably low. Therefore no study of LPC improvement would be complete without presenting a satisfactory approach designed to minimize acoustic-noise interference. We have already presented other interrelated factors which result in speech degradation. We next present a preferred solution to the problem of LPC processing of noisy speech.

Factors Critical to LPC Performance

Extensive test data collected by DoD agencies have shown that noise suppression alone does little to improve the quality of noisy LPC speech. Other areas, such as speech enhancement and proper LPC analysis and synthesis, should be dealt with in combination with the noise suppression. All the topics we have discussed are integral parts of the solution to the problem of noisy speech, but there are other factors as well. The talker must speak carefully and hold the microphone properly. Care must also be taken to ensure that the audio circuits and microphones used do not introduce any anomolies of their own which would be detrimental to the LPC analysis. Above all, the noise suppression must be based on realistic assumptions about noise characteristics. We now discuss each of these other factors.

Talker Performance

People do not talk the same way in the presence of intense ambient noise as they do in a quiet environment. They tend to raise their pitch and to speak louder in an effort to overcome the background noise [16,17]. In a vehicular platform, such as a tank, jeep, or armored personnel carrier, the talker's chest and throat are subjected to vibrations that cause the voice to shake.

Efforts have been made to optimize the performance of talkers—cause them to speak slower, articulate more carefully, or hold the microphone properly—by manipulating the sidetone [18]. Results indicate that sidetone delays of up to 30 ms do cause talkers to speak more slowly but that this slight change in speaking rate is not enough to significantly improve the quality or intelligibility of the processed speech.

Actual noisy speech cannot be approximated by computer-generated speech in which noise-free high-quality speech is combined either with actual noise or with computer-generated noise. Therefore a noise-suppression method optimized using this type of simulation may not work well in the field.

Noise Level

Ambient noise levels vary widely. In an office environment the noise level can be as low as 60 dB, causing virtually no interference with the speech; in a helicopter or a tank it can be as high as 115 dB. Since the normal speaking level as measured about 6 or 7 mm (1/4 in.) from the mouth is between 105 and 115 dB, the noise level in some platforms will be louder than the speech at the microphone. The narrow LPC begins to show perceptible degradation with a noise level of approximately 80 dB.

According to extensive data collected by the Navy [19], intelligibility of the narrowband LPC depends more on the noise level (the speech-to-noise level for a given speaking level) than on noise characteristics. We generated Fig. 13 by using the DRT data in Ref. 19 and the actual speech-to-noise ratios computed from audio recordings made in the platforms. The figure shows the approximately linear relationship between the intelligibility and the speech-to-noise ratio. Thus, some form of noise reduction is essential for improving the narrowband LPC performance in the presence of noise interference. The overall effect, however, may not be significant if the noise suppression is accompanied by speech distortions, as we will discuss later.

In the past much emphasis has been placed on electronic and digital processing approaches to noise reduction, and mechanical approaches have been largely ignored. Speech and background noise do not originate at the same point, so in some platforms it may be feasible to acoustically insulate the speaker or microphone from at least some of the noise. Since it is easier to simply prevent the noise from contaminating the signal than to remove the noise afterward, such possibilities should not be overlooked. Another advantage of acoustic insulation is that, unlike virtually all noise-suppression devices, it does not introduce speech distortions.

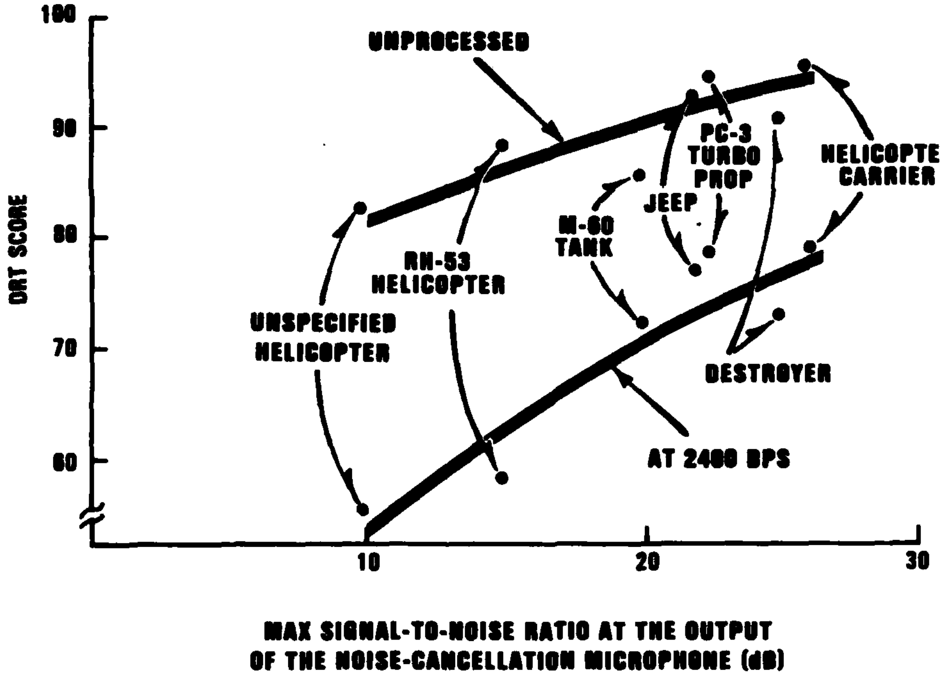


Fig. 13 — Intelligibility of the narrowband LPC in terms of the speech-to-noise ratio

Noise Characteristics

All current noise suppressors with a single audio input are based on the assumption that noise is stationary. This assumption is one reason they do not perform satisfactorily in the field. Noise is rarely stationary, even when the platform maintains a steady motion. Figure 14 shows a superposition of 20 amplitude spectra of helicopter noise, each spectrum being for a separate 22.5-ms frame. It can be seen from the spread of the 20 curves that spectral deviations from frame to frame are substantial.

Actually much platform noise is even less stationary than that in this example. Gunshots may be heard sporadically. Tank noise changes noticeably as the platform accelerates. Helicopter noise has a rotor-modulation component with frequencies lower than that of the LPC frame rate. Communication centers are usually jammed with on-and-off tones from other equipment and with interfering human voices.

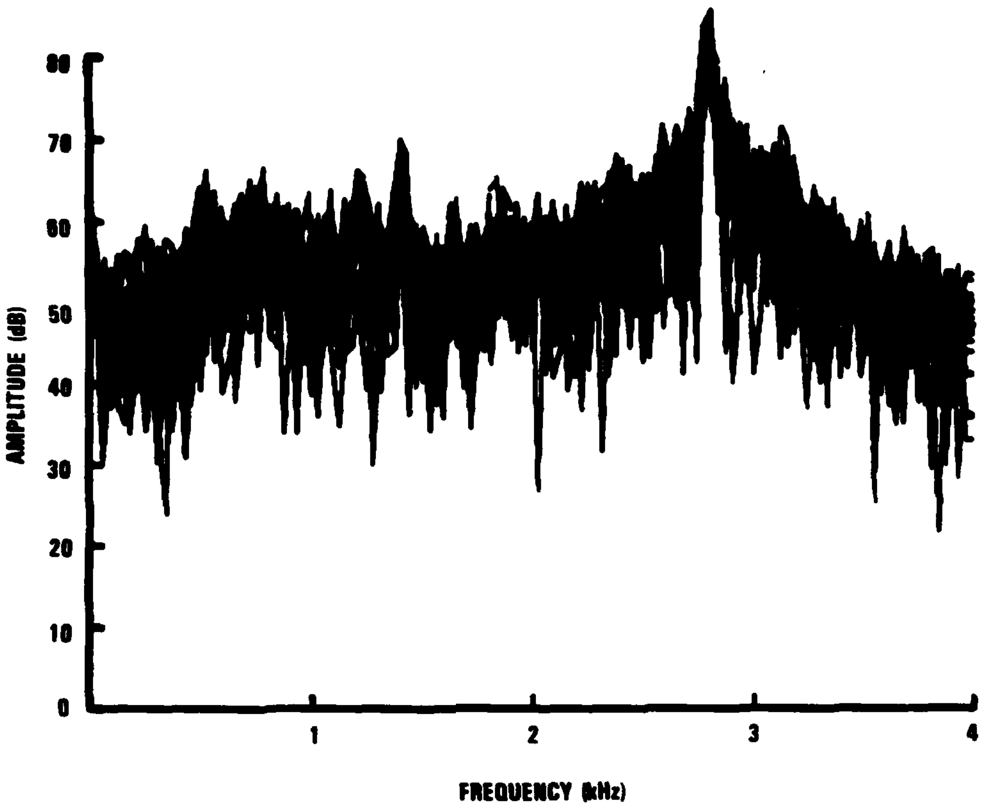


Fig. 14 — Overlay of 20 spectra for 20 frames of RH-53 helicopter noise

Noise-suppression preprocessors having a single audio input estimate noise parameters—correlation coefficients, frequency components, etc.—during silent periods and update them whenever there is no speech activity. These estimated parameters are subsequently used for the noise reduction. Parameter variations from frame to frame reduce the efficiency of the noisy reduction, but a more serious problem is the reliable discrimination of nonspeech from speech. For example, if fricative waveforms are erroneously detected as ambient noise, the noise suppressor will consequently reduce fricative sounds at the output.

Noise-Cancellation Microphones

A noise-cancellation microphone is an acoustic transducer which has a built-in noise-suppression feature [20]. It may be operated alone or in tandem with a noise-suppression preprocessor, depending on the severity of the noise environment. All noise-cancellation microphones currently deployed are first-order gradient microphones which measure the pressure difference between two closely spaced points. The pressure at each point is caused by both speech and noise. (This is distinctly different from a noise suppressor with two separate microphones.) The pressure difference is a function of the frequency, distance, and arrival angle of the sound. The near-field response is independent of frequency, whereas the far-field response is a linear function of frequency, analogous to a filter having a frequency-dependent gain of 6 dB per octave [20]. The noise-cancellation microphone cannot distinguish speech from noise, so it rejects incoming sounds based on the distance to the sound source and the arrival direction of the signal. Hence it is essential that the microphone be held close to the mouth, not only to have a proper output level but also to minimize the attenuation of low-frequency speech components.

The use of a well-designed noise-cancellation microphone is a simple means of noise suppression, particularly when the noise contains predominantly low-frequency components. As Fig. 13 shows, the intelligibility of unprocessed speech from a noise-cancellation microphone is adequate for voice communications. Apparently the human auditory system is better able to filter out background noise if the low-frequency interference has been removed. Unfortunately the LPC analysis does not have such perceptive powers—it simply tries to model the speech spectral envelope with ten poles. Thus the presence of high-frequency noise not filtered out by the noise-cancellation microphone will produce spectral estimation errors in both high- and low-frequency regions.

Ideally the LPC analysis requires a noise level at least 15 dB below the speech spectral envelope at all frequencies, not just at the lower frequencies. Actually noise reduction is more important for higher frequencies than for lower ones, since the speech spectral envelope is generally weaker in the upper regions. Likewise noise suppression is needed more for the interformant regions than around formant frequencies. At present the use of a noise-cancellation microphone in a tank platform gives a speech-intelligibility (DRT) score in the low 70s, considered barely acceptable, when processed through the narrowband LPC.

Spectral Distortions Prior to LPC Analysis

Not all distortions in the speech spectrum are caused by ambient noise; most microphones and noise-suppression preprocessors also introduce distortions. In certain cases these distortions are even more severe than those from ambient noise. If a noise-reduction process merely trades one form of distortion for another, the narrowband LPC performance will not improve as much as might otherwise be expected. We examine some of these distortions.

All noise-cancellation microphones currently deployed were originally designed for analog speech transmission in which a flat amplitude response was not really essential. A microphone can have severe distortions in its amplitude response and still provide a usable voice signal because of the great redundancy in the speech waveform and the ability of the human ear to compensate for such distortions. Since the LPC analysis performs spectral analysis with a limited number of poles, any false frequencies created by microphone resonance will introduce spectral estimation errors. Figure 15 shows the amplitude response of a typical noise-cancellation microphone (the M-87). Ripples are obvious in the passband. The use of such a microphone is one reason noise-suppression preprocessors have shown little promise in field tests.

The amplitude-response curve of a microphone indicates only the steady-state behavior under the excitation of a discrete frequency at a constant power level. It does not reveal the transient performance under the excitation of a broadband signal for a wide dynamic range. A microphone with a flat amplitude response may still generate considerable distortions where the waveform changes abruptly, as it does at onsets or at the beginning of a pitch cycle. The latest experimental noise-cancellation microphone, used to produce the spectrogram in Fig. 16, has a nearly flat amplitude response within the passband but gives distortions similar to those of a carbon microphone (Fig. 1b). The strong spreading of the fricative spectra, such as /f/ and /th/ in Fig. 16, causes them to sound more likes bursts than like fricatives. As a result "Fred" becomes "pred", and "think" sounds like "tink". The words "does" and "much" are identifiable, but the /z/ and /ch/ sounds are noticeably distorted. Burst spectra are also affected, causing the /t/ in "too" to be grossly overemphasized. Consequently the DRT score for noise-free speech from this microphone is only 78 at the output of the narrowband LPC.

Noise-suppression preprocessors can also generate distortions with noise-free speech. Failure of the transfer function of the preprocessor to reduce to a unity gain in the absence of noise is a good indication that the noise suppressor will introduce speech distortions. In the past, speech intelligibility has actually been degraded through the use of certain noise-suppression preprocessors, with or without the narrowband LPC in tandem and with or without additive noise.

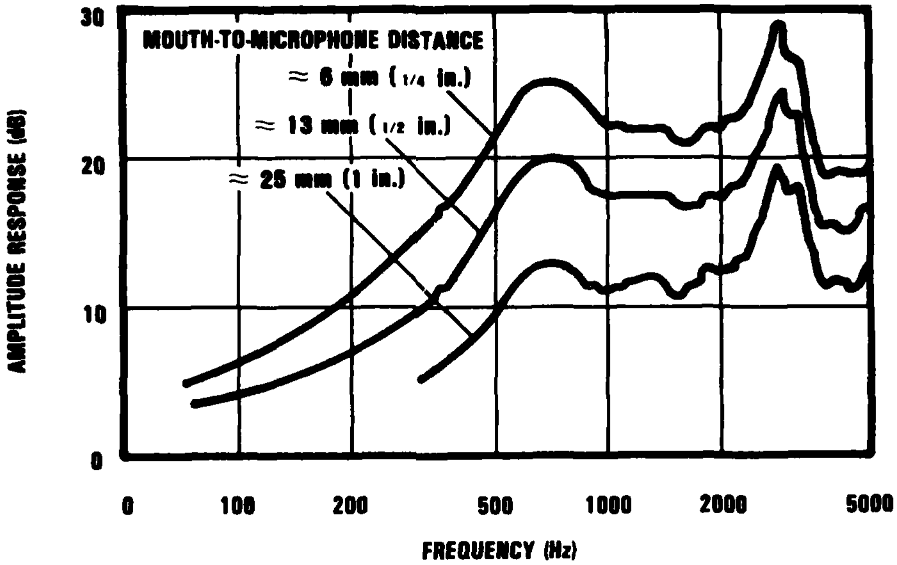


Fig. 15 — Amplitude response of an M-87 noise-cancellation microphone

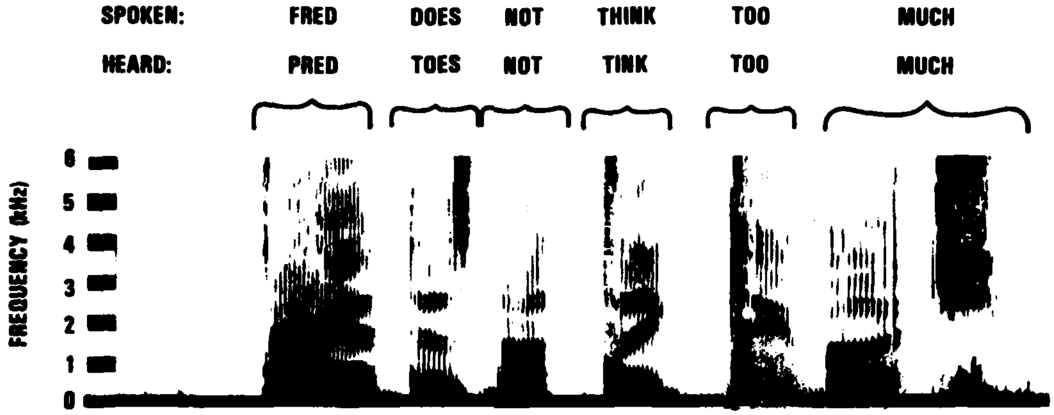


Fig. 16 — Spectrogram of a distorted output from a microphone

It is important for the noise suppression device, whether it is a microphone, a noise-suppression preprocessor, or a combination of both, to provide good speech intelligibility in the absence of noise interference. This is because a noisy platform, such as a tank or a jeep, may turn into a quiet platform if the engine is turned off. In a tactical situation voice communications might still take place.

Automatic Gain Control

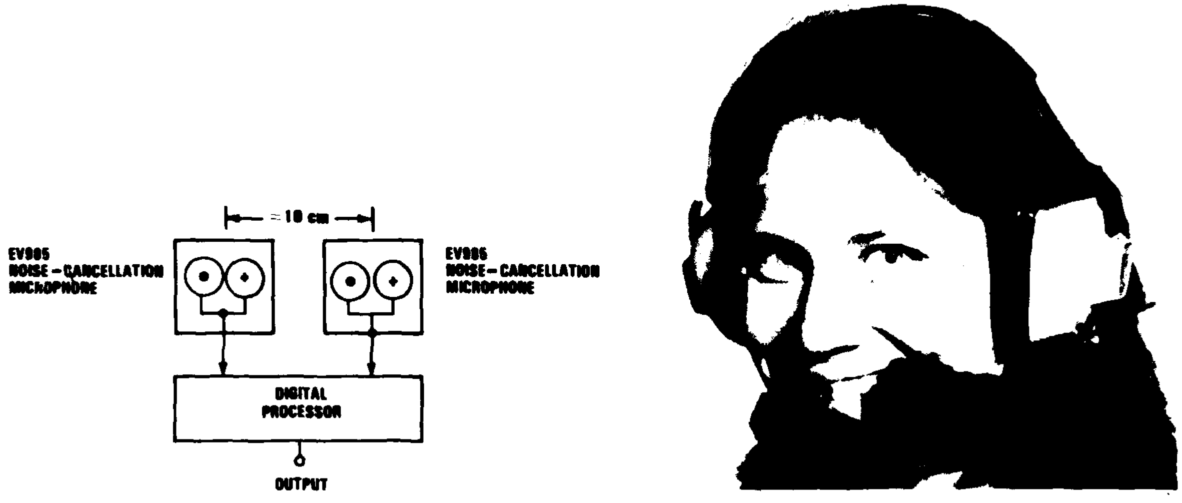
A noise-suppression preprocessor with a single audio input may require a time-invariant front-end audio gain. This is true if the noise suppressor estimates amplitude-dependent noise parameters during intervals of silence. If the audio gain changes during voiced periods, though, the noise parameters are not longer valid.

The design of a noise-suppression preprocessor must take into account the availability of a constant audio gain in the front end. Some analog circuits deployed in platforms may have automatic-control circuits to benefit the analog speech transmission, and some of the current DoD narrowband LPCs have an automatic-gain-adjustment feature.

Our Noise-Suppression Preprocessor

As evidenced by numerous articles in the conference records of the IEEE International Conference on Acoustics, Speech, and Signal Processing over the past several years, researchers have tried many forms of noise suppression. DoD agencies have also sponsored a number of programs with the same objectives and have selected a few approaches for extensive testing. So far the results have not been too encouraging. As we just discussed, no technique of noise-suppression preprocessing will significantly improve the narrowband-LPC performance if it has to operate with distorted speech or if it is based on the unrealistic assumption that noise is stationary.

Therefore we propose a different form of noise-suppression preprocessor. In essence the device is a hybrid acoustic transducer which combines a digital signal processor and two carefully selected noise-cancellation microphones 10 cm apart. Since the two microphones are placed side by side in the conventional boom-microphone configuration, there are no installation problems, and there are no differences in handling on the part of the operator (Fig. 17). The two major advantages of this approach are that our noise preprocessor has its own microphones and requires no others and that our noise-suppression algorithm is not based on the assumption of stationary noise.



(a) Functional diagram

(b) Laboratory setup

Fig. 17 — Our noise-suppression preprocessor

The microphone we selected, the Electrovoice EV985, is best available noise-cancellation microphone. It not only has a flat amplitude response within the passband but also has an excellent transient response at onsets. The microphone about 10 cm from the mouth picks up virtually no speech, so it provides information on noise alone, with the noise not being far different from the noise picked up along with speech by the microphone about 6 mm from the mouth. Another advantage of this noise-suppression preprocessor is that it does not try to estimate noise parameters during silent periods.

The digital processing of the microphone signal is an important consideration in the design of any noise-suppression preprocessor. The processing can be relatively simple or complex, depending on the ultimate objective of the noise suppression. In the present application the objective is to improve the narrowband LPC. Therefore the noise suppressor is required only to refine speech parameters as defined by the LPC and not necessarily to refine the speech waveform in general or its amplitude and phase spectra. Since the LPC performs a correlation analysis, the phase spectral information is not as critical as the amplitude spectral information. Therefore only the amplitude spectrum of the speech plus noise is adjusted by the amplitude spectrum of the noise alone.

The noise source in a military environment is rarely a single point-source far from the speaker. This is particularly true with extremely noisy platforms in which noise suppression is most needed, such as tanks, helicopters, and armored personnel carriers. In these platforms the communicator is in a small compartment in which the surrounding walls are literally the noise sources. Since the radiation characteristics of a noise source vary with its frequency content, the phase difference between the two microphones also changes rapidly with the nonstationary noise encountered in these platforms. (Fluctuations of the phase difference can be readily observed from the Lissajous pattern.) Thus noise suppression based on amplitude and phase compensations, such as the algorithm advanced by Widrow et al. [21], may not be highly effective. However, this approach is under further investigation by some of our coworkers.

Unfortunately this approach, like other two-microphone approaches [22], needs a microphone separation of several meters. But it is practically impossible to install the distant second microphone in crowded platforms such as tanks or helicopters. Our noise suppressor does not present this installation problem.

Our Algorithm

In our algorithm noise suppression is performed once per frame. The frame size could be identical to that of the LPC analysis if the suppressor output were to be fed directly into the LPC. The output of each microphone is partially overlapped with the trailing samples of the preceding frame in order to suppress discontinuities at frame boundaries and to suppress spectral leakage. Since the outputs of both channels are real, a single complex fast-Fourier transform may be used to generate both frequency spectra.

The principle of the noise reduction is essentially identical to that of the "spectral subtraction method" [23]. Let the time-overlapped noisy speech sample be represented by

$$x(i) = s(i) + n(i), \qquad (13)$$

where $s(i)$ is the speech sample and $n(i)$ is the noise sample. If the speech sample is uncorrelated with the noise sample, the power spectrum of the noisy speech is the sum of the individual power spectra. Thus

$$|X(k)|^2 = |S(k)|^2 + |N(k)|^2, \qquad (14)$$

where $X(k)$ is the amplitude spectrum of the noisy speech, $S(k)$ is the amplitude spectrum of the noise-free speech, and $N(k)$ is the amplitude spectrum of the noise from the speech channel. Although $N(k)$ is not available, the amplitude spectrum of the noise from the speech-free channel is, and we have assumed that the amplitude spectra (not the phase spectra) of both noises are similar, since the microphones are purposely placed only 10 cm apart. Hence $N(k)$ may be closely approximated by $\hat{N}(k)$, the amplitude spectrum of the noise from the speech-free channel.

The complex frequency spectrum of noise-suppressed speech with the original phase spectrum (which has been corrupted) is

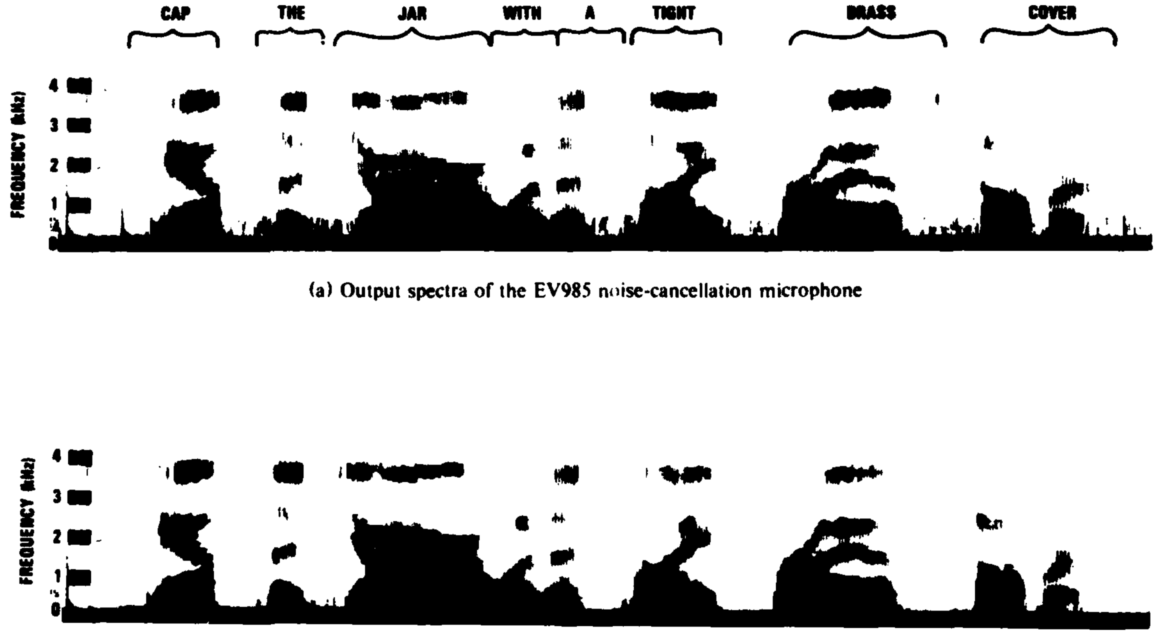
$$\hat{S}(k) = X(k)\sqrt{1 - \frac{|N(k)|^2}{|X(k)|^2}}. \tag{15}$$

In the absence of noise the transfer function of the noise suppressor is a unity gain; hence there will be no inherent speech distortion in a noise-free environment. If the quantity under the radical is negative, it should be set to zero or, preferably, to a fraction of the actual amplitude spectral value (−12 dB or so). According to informal tests the presence of a nonzero spectral floor gives rise to more natural sounding noise-suppressed speech.

The noise-reduced speech spectrum is converted to time samples by the inverse fast-Fourier transform. Time-overlapping of each frame's samples with the trailing samples of the preceding frame, which must be consistent with the time-overlapping prior to the fast-Fourier transform, gives the noise-suppressed speech samples.

Test Results

Using our noise suppressor, one male and one female speaker recorded six sentences each in the presence of helicopter noise. The sound pressure level was 115 dB, the level found in the actual platform. The spectrogram of the output of the EV985 noise-cancellation microphone (Fig. 18a) shows a fair amount of noise still present in the background. Figure 18b shows that our noise suppressor removes much of the background noise without introducing any apparent distortions in the speech spectra.

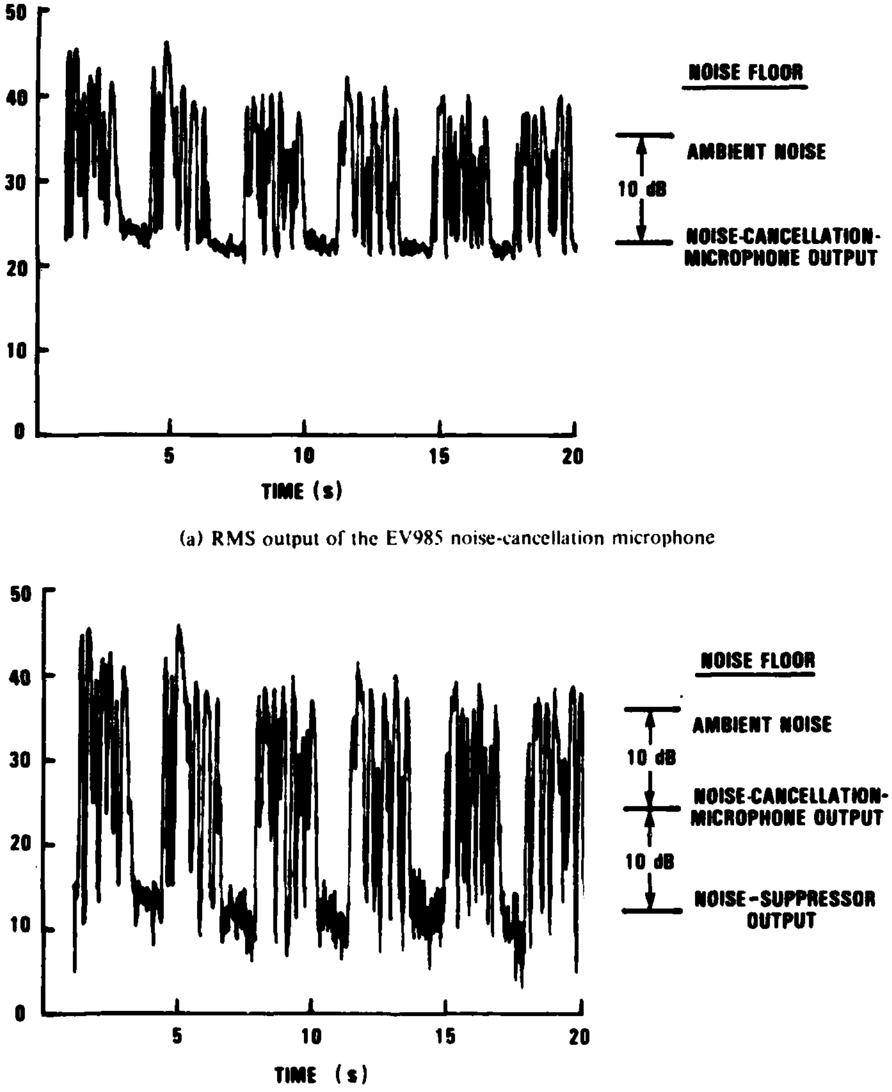


(a) Output spectra of the EV985 noise-cancellation microphone

(b) Output spectra of our noise suppressor

Fig. 18 — Performance of our noise suppressor

A quantitative measure of the noise reduction is shown in Fig. 19, which illustrates the envelope of the short-term averaged speech RMS values from an EV985 noise-cancellation microphone. The background noise is helicopter noise at a sound pressure level of 115 dB. Each RMS value is based on 250 nonoverlapping samples. The figure shows the RMS envelope of five 3-s sentences with a 1-s pause between sentences. As noted, the use of the noise-cancellation microphone alone reduces the noise floor by approximately 10 dB. Figure 19b is a similar plot of the output of our noise suppressor. The noise floor is further reduced by another 10 dB. In this example the use of our noise suppressor improves the dynamic range of the speech by 20 dB, which is a significant improvement.



(a) RMS output of the EV985 noise-cancellation microphone

(b) RMS output of our noise suppressor

Fig. 19 — Noise levels

Summary

Degradation of LPC performance under conditions of acoustic noise interference is caused by a multitude of factors, including talker performance, noise level, noise characteristics, microphone characteristics, spectral distortions prior to the LPC analysis not caused by noise, and improper LPC analysis (erroneous window placement at onsets). To improve the LPC performance in a noisy environment, these factors must be examined comprehensively.

Many previous noise suppressors have shown little promise because:

- The noise is assumed to be stationary, which is rarely the case in a military environment,
- Noise characteristics are estimated during silent periods, although there is no reliable way of detecting these periods when the background noise is as loud as the speech, and
- The noise suppressors often introduce speech distortions and do not exhibit a unity gain *in the absence of noise.*

Our noise suppressor does not suffer from these limitations. It exploits the best features of a noise-cancellation microphone—good noise suppression at low frequencies and good speech isolation at a distance of several centimeters, and it compensates for the lack of noise cancellation at high frequencies with the digital signal-processing techniques. An important aspect of this noise suppressor is the use of two carefully selected high-performance noise-cancellation microphones (Electrovoice 985), which are housed in the standard boom-microphone configuration. Most significantly the computational complexity of this approach is of the same order of magnitude as that for a noise suppressor using a single microphone and the identical noise-suppression principle.

AUTOMATIC GAIN CONTROL

The purpose of an automatic gain control (AGC) is to self-adjust the front-end gain of the LPC analyzer in such a way that the speech waveform is more accurately quantized by the analog-to-digital converter. Tests in the past have indicated that properly amplified speech produces higher intelligibility scores at the narrowband LPC output because both filter and excitation parameters are more accurately estimated. In addition, properly amplified input speech results in properly amplified speech at the receiver, which is highly desirable for listening in a noisy environment.

Need for AGC

If the narrowband LPC were always operated with a well-matched single audio input, an AGC would not be needed. In a military environment, however, such an ideal setup cannot be guaranteed. For instance the narrowband LPC might be connected to existing intercom systems, which would have different gains from one platform to another. Likewise the LPC might be operated in tandem with another type of voice processor having a different output level.

The LPC front end could be equipped with a manual gain control to compensate for the external gain mismatch. It has been reported, however, that the manual gain control in the previous voice processors did not work well, because the operators in the field often did not know how to adjust the gain properly. Thus an automatic gain control is the preferred solution.

Background

Not all AGC devices are suitable for use with the narrowband LPC. For example, a fast-attack-and-slow-release AGC can be dismissed from further consideration, since this kind of AGC is basically designed for unprocessed or waveform coders such as continuously-variable-slope delta (CVSD) modulators not involving analysis and synthesis. Since LPC parameters are derived from a short-term correlation of the speech waveform, any time-varying amplitude weighting caused by a gain change within the analysis window will introduce errors in the estimated parameters. The AGC gain is expected to change most radically at speech onsets, which contain much critical information. Hence a slowly adjusting gain device is preferable for compensating the external gain mismatch.

A major difficulty in estimating the necessary gain is that the speech envelope varies continuously even if the external gain is properly matched. The magnitude of the speech envelope not only depends on the external gain and the speaker's efforts to control loudness (the volume of air from the lungs) but also depends heavily on the resonant characteristics of the vocal tract. Thus any usable gain-control signal must be somewhat independent of the formant distribution.

Our AGC Implementation

The performance goals for which we implemented an AGC are as follows:

- The AGC function shall not introduce speech delay;

- The computation time shall be a few percent of that required for the LPC analysis and synthesis;

- The DRT score shall not decrease as the external gain is changed from 0 to −28 dB;

- The gain level shall be stabilized within 5 s of the initial onset of voiced frames;

- Once a stable gain is reached, there shall be no noticeable haunting (undesirable undulation of loudness) for normal speech;

- There shall be no gain pumping during nonvoiced periods, even if the speech is corrupted by severe noise such as helicopter noise.

In our AGC approach the necessary gain is estimated by digital computations. The estimated gain is then fed back to the analog amplifier at the LPC front end, as shown in Fig. 20. Our approach is based on the feedback control principle, whereby the quantity derived from the gain-adjusted speech waveform (which we will define later) is compared with the reference level. The resulting difference, or error, is the control signal that adjusts the front-end analog amplifier gain. If the input is unvoiced (consonants) or nonvoiced (not speech), the gain will not be computed or adjusted.

The control signal is generated by the use of the speech energy contained in a low-frequency band below 1 kHz. The magnitude of this low-band energy—primarily the first-formant amplitude—is relatively independent of the nature of the spoken sound. Another advantage of using the low-band energy is that it is available in most LPCs as a byproduct of the voicing decision.

Each low-band energy is a short-term averaged quantity. Therefore it is susceptible to both internal and external perturbations such as loudness fluctuations, leakage of higher formant frequency components, or ambient-noise interference. To smooth out such fluctuations, the mean value of the low-

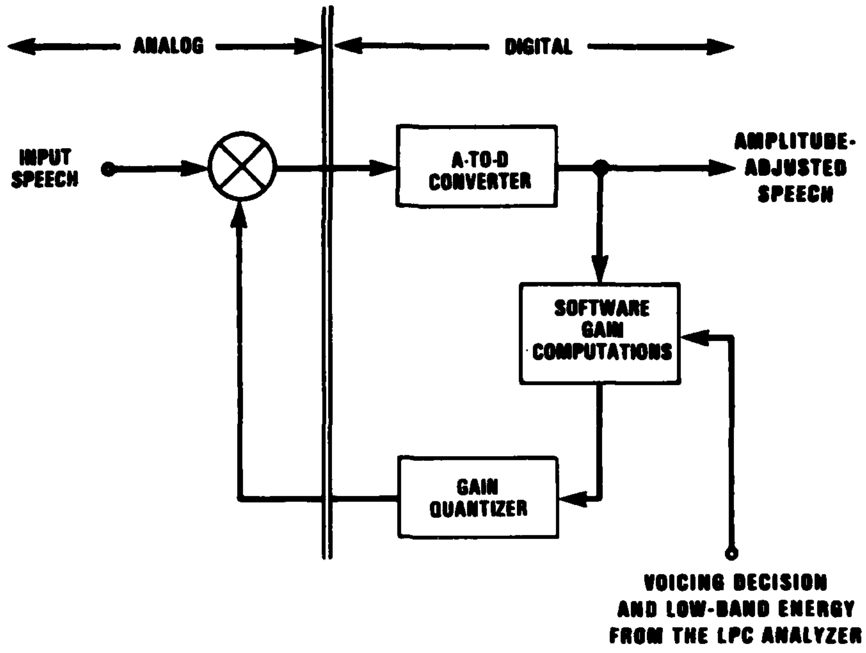

ANALOG
DIGITAL
INPUT SPEECH
A-TO-D CONVERTER
AMPLITUDE-ADJUSTED SPEECH
SOFTWARE GAIN COMPUTATIONS
GAIN QUANTIZER
VOICING DECISION AND LOW-BAND ENERGY FROM THE LPC ANALYZER

Fig. 20 — Our automatic gain control

band energy is computed by the use of the probability density function of the low-band energy. The probability density function is computed from the low-band energies for the past several seconds (voiced frames only) and is continuously updated at each voiced frame.

To facilitate computations, each low-band energy is quantized from −20 dB to 20 dB around the reference level in equal-decibel steps, and the probability density function is computed for those input values. We found that a fixed step size of 1.75 dB is acceptable, giving 25 input values (Table 5). Since the AGC seeks a null-error condition, neither the size of the input step nor the range is too critical to the overall AGC performance.

Despite the computational complexities, the use of the statistical average is preferred to the time average, because low-band energy is not uniformly distributed. (If it were, both results would be identical.) In addition, the statistical average is less susceptible to sporadic interference, because the average has been weighted by the duration of the event. The estimated mean is remarkably constant for normal speech, indicating that it is relatively free from the characteristics of vocal-tract resonance.

The difference between the reference level and the estimated mean of the low-band energy controls the front-end analog amplifier gain. The incremental gain and the error are related by a transfer characteristic which is linear except for a dead zone near the zero-error region. The rate of gain update may be changed by altering the slope of the linear region.

The reference level is established by the analysis of the low-band energies of properly amplified voiced speech waveforms. Since a 12-bit analog-to-digital conversion is standard, the maximum speech amplitude without clipping is between −2047 and +2048. The analysis indicates that low-band energies of voiced speech waveforms are clustered around 250. This is the assigned reference level.

Table 5 — Characteristics of the Low-Band Energy Quantizer

Quantized Level (Y)	Low-Band Energy* (y_i)	Quantized Level (Y)	Low-Band Energy (y_i)
1	22 or less	14	306
2	27	15	374
3	33	16	458
4	41	17	560
5	50	18	685
6	61	19	837
7	75	20	1024
8	92	21	1253
9	112	22	1534
10	137	23	1870
11	167	24	2298
12	205	25	2813 or more
13	250†		

*Based on 12-bit input speech.
†Reference level.

Our AGC Algorithm

Figure 21 is the flow chart of the AGC gain computations. As indicated, the entire operation is bypassed if the speech is unvoiced. The low-band energy is quantized to one of 25 1.75-dB steps, as we just discussed. Thus

$$y_i = F(x_i), \tag{16}$$

where x_i and y_i are low-band energies before and after quantization respectively and $F(\ldots)$ is the quantization operation based on the rule shown in Table 4.

To compute the probability density function of the quantized low-band energy, one register is assigned to each quantization level. When the quantized low-band energy is equal to a particular counter index, the content of that register is incremented by one. The contents of all registers are then short-term averaged by a single-pole filter with a feedback constant of 1/32. Thus

$$C_i(Y) = C_{i-1}(Y) + [A_i(Y) - C_{i-1}(Y)]/32, \tag{17}$$

where $C_i(Y)$ is the content of the register associated with quantization level Y during the ith voiced frame. The incremental count $A_i(Y)$ is expressed by

$$A_i(Y) = \begin{cases} 1, & \text{if } y_i = Y, \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

The feedback constant in Eq. (17) defines the width of an exponentially decaying window for the registers. According to tests with a real-time simulator using a variety of speech samples, a feedback constant of 1/32 is a suitable choice in terms of fast gain settling without introducing undesirable haunting in the steady state. The feedback constant 1/32 in Eq. (17) does not directly control the gain update rate; the gain updating factor is an incremental gain Δg_i, which we now define.

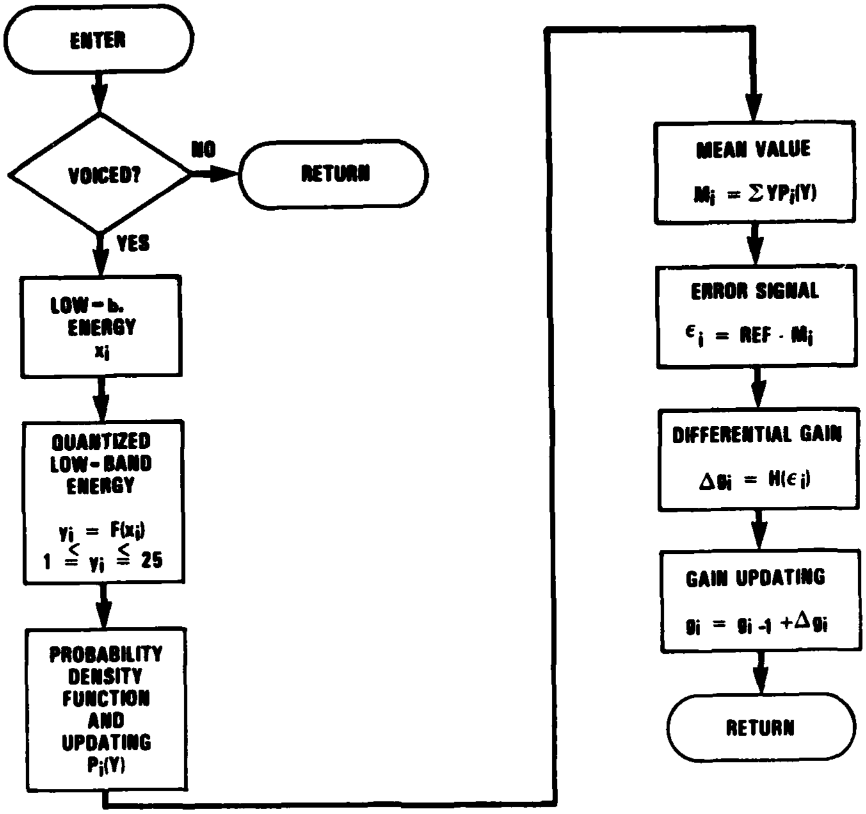


Fig. 21 — Gain computations in our automatic gain control (Fig. 20)

Using updated register contents, the probability function is computed by

$$P_i(Y) = \frac{C_i(Y)}{\sum_{Y=1}^{25} C_i(Y)}. \tag{19}$$

The error is defined as the difference between the reference level (*REF* in Fig. 21) and the mean of the low-band energy. Thus

$$\epsilon_i = REF - \sum_{Y=1}^{25} YP_i(Y). \tag{20}$$

As we noted in Table 5, the reference level is 13.

The front-end analog amplifier gain in decibels, as denoted by g_i, is incrementally adjusted by

$$g_i = g_{i-1} + \Delta g_i. \tag{21}$$

where the incremental gain Δg_i in decibels is nonlinearly related to the error:

$$\Delta g_i = \begin{cases} 0, & \text{if } |\epsilon_i| \leq 2, \\ (\epsilon_i + 2)/32, & \text{if } \epsilon_i < -2, \\ -(\epsilon_i - 2)/32, & \text{if } \epsilon_i > 2. \end{cases} \tag{22}$$

The transform characteristic has a dead zone near the reference level and is linear elsewhere. Thus, if the estimated mean of the low-band energy is within two quantization levels (3.5 dB) of the reference level, no gain adjustment will be made.

There is a broad range of acceptable update factors. We chose the factor 1/32 after experimenting with various types of speech input, including noisy speech and lengthy two-way casual conversations over a real-time processor. Our decision was based on both the transient and the steady-state performance, in particular on the gain settling time from an initial gain mismatch as large as −28 dB and on the amount of haunting once the final gain is reached.

Test Results

Our AGC function has been incorporated in the NRL-owned programmable LPC processor and in one of the DoD narrowband LPCs currently under development. We compared performance with the initial design goals we listed previously:

- The AGC established the necessary gain based on past speech statistics. Therefore no additional processing delay is introduced beyond that required for the narrowband LPC.

- With use of assembly language, the computation time was 0.55 ms for voiced frames and 0.015 ms for unvoiced frames. If microassembly language is used (which will be the case in the final implementation), these computation times may be reduced.

- Scores of a single-speaker DRT are virtually unchanged as the external gain mismatch is changed from 0 dB to −28 dB.

- The transient response of the AGC is remarkable. The steady-state gain is reached within a few seconds even if the input gain mismatch is as much as −28 dB.

- Once the steady state is reached, there is no noticeable haunting. This conclusion is based on a recent 30-min recording of two-way conversations of various speakers through the narrowband LPC with the AGC in the front end.

- If the narrowband LPC were to make voicing errors (from "unvoiced" to "voiced"), the AGC gain would be increased steadily during unvoiced periods. Such voicing errors did not happen, even with helicopter speech samples, but there is the possibility of gain pumping in other unexpected circumstances. Therefore we also present a method of noise suppression and an improved pitch and voicing estimator.

REMARKS ON THE ANALOG CIRCUITS, THE MICROPHONE SHIELD, AND PITCH AND VOICING ESTIMATION

We have discussed five new areas we investigated for improved LPC analysis. We now discuss some more well-known areas which still need improvements: the front-end analog circuits, the effects of the microphone shield, and the pitch and voicing estimation. Though we discuss each of these areas only briefly, they are important.

Front-End Analog Circuits

In one case during the early days of narrowband-LPC development, a complete redesign of the front-end analog circuits improved the DRT scores 5 points, which illustrates the critical nature of the analog circuits. Four important aspects of analog circuits are the following:

- The analog circuits need to be completely isolated from the digital circuits. Leakage of clock pulses into the analog circuits must be totally suppressed, and the internally generated noise cannot exceed the least significant bit at a 12-bit analog-to-digital converter output in order to ensure a 60-dB dynamic range for the speech signal.

- Speech should be bandpass-filtered, rather than lowpass-filtered, prior to the analog-to-digital conversion. Attenuation of extreme-low-frequency components is beneficial, because any disturbance in that frequency region creates an interference of a wider bandwidth at the output of the narrowband LPC. This is due to the insufficient frequency-resolution capability of the filter weights, particularly during the silent periods, when the DoD narrowband LPC transmits only four coefficients. The preferred low-frequency cutoff point is 150 Hz, with an attenuation rate of 18 dB per octave. When the cutoff frequency is raised to 250 Hz, the DRT score is reduced by 3 points. A cutoff frequency of 250 Hz at the receiver does not affect the DRT score.

- The upper cutoff frequency of the bandpass filter is also a significant parameter. At one time there was a narrowband LPC that used an upper cutoff frequency of 3.2 kHz, but the speech intelligibility of this processor was in general lower than that of other narrowband LPCs which used a higher cutoff frequency. Currently the DoD narrowband LPC uses an upper cutoff frequency of 3.6 kHz, but there is no unanimity on this choice. To determine the preferred upper cutoff frequency, we constructed an analog bandpass filter with an adjustable upper cutoff frequency. The filter has a fixed attenuation rate of 100 dB per octave. As the upper cutoff frequency is changed from 3.6 kHz to 4 kHz, the LPC processed speech becomes decidely brignter, and stop consonants sound sharper. The DRT score (three male speakers) with the 3.6-kHz cutoff is 83; the score with the 4-kHz cutoff is 87. With three female speakers, the scores are 79 and 83 respectively.

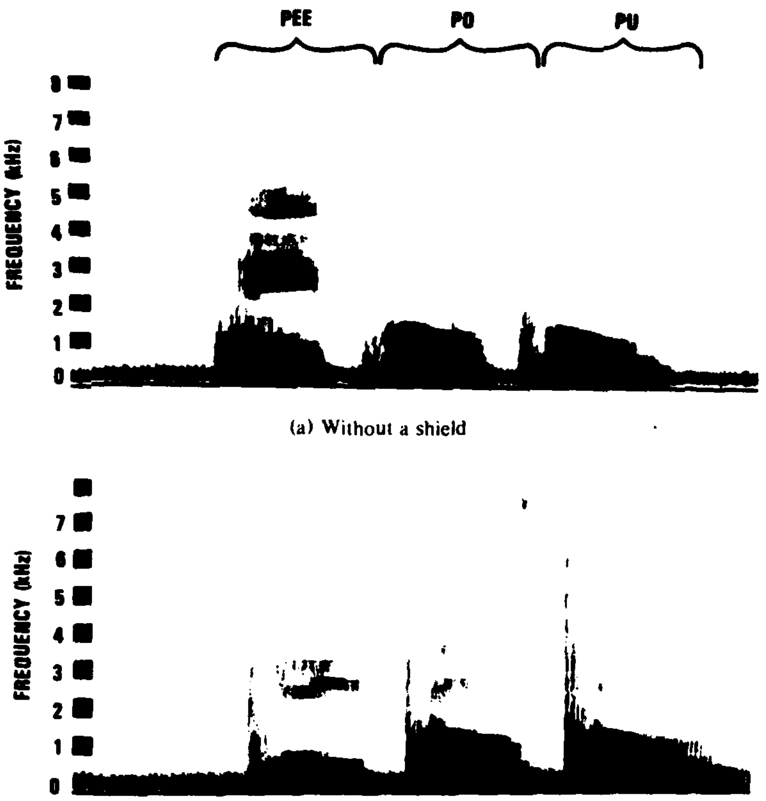
- The amplitude response of the analog filter within the passband must be flat. This requirement is readily understandable, because the front-end analog circuits should not introduce false formant frequencies. The requirement of the phase response, however, has never been well defined. Distortions of the phase response create reverberant speech sounds, and experimentation indicates that reverberant speech is not well reproduced by the narrowband LPC. The best approach to achieving a front-end frequency response with a sharp cutoff characteristic and a linear phase response is through the use of a digital filter at double the sampling rate, (such as the special-mode filter in Fig. 2 and Table 1).

Effect of the Microphone Shield

The quality of LPC-processed speech depends directly on the quality of the microphone used. As we discussed previously in connection with the improvement of LPC speech in noise, a flat amplitude response and a good transient response over a wide input dynamic range are indispensable requirements for any microphone which is to be used with the narrowband LPC.

One aspect of the microphone—the effect of the microphone shield on the LPC processed speech—has been neglected in the past. The microphone shield in a typical telephone handset is plastic pierced with a circular array of small holes. This creates an irregular surface which alternately reflects and passes the incoming sound waves. As a result, speech sounds, particularly plosives such as /p/, are spattered on impact, causing the spectrum to spread noticeably (Fig. 22). The spectrum of /p/ normally has predominantly low-frequency components; a burst with relatively strong high-frequency components is characteristic of /t/. Therefore the spattering effect of the microphone shield causes /p/ to sound more like /t/.

The narrowband LPC tends to accentuate these distorted plosive sounds. They are frequently heard as pops, and the voicing decision tends to be "voiced" rather than "unvoiced" as it should be. Not only is the intelligibility of the plosive sound itself reduced, but conflicting burst and vowel transition information (such as a /t/-like burst followed by formant movements typical of /p/) can be confusing to the listener.



(a) Without a shield

(b) With the standard shield used in a telephone

Fig. 22 — Effect of a microphone shield on the speech spectrum of /p/

Figure 23 shows a microphone assembly we designed. We used this microphone in experiments with selective aliasing at the front-end of the LPC, as we discussed earlier. The assembly consists of a commercially available electret microphone element, with the pierced section of the conventional microphone shield replaced by soft foam. The foam insulates the microphone element from strong puffs of air but does not interfere with the passage of the sound waves as does the standard plastic shield. This assembly is ideal for laboratory work, but it is not rugged enough for general use; the microphone element should be further protected by a wide-grid screen over the foam.

Pitch and Voicing Estimation

Pitch and voicing estimation has always been the most difficult operation in any narrowband voice processor. Recently, the use of high-speed digital signal processors has made possible more complex data processing, elaborate logical operations, and delay decisions. The dynamic programming approach to pitch tracking (called DYPTRACK) advanced by William Blankenship [24] has also been a significant contribution. As a result pitch and voicing estimation has been improved markedly in the current narrowband LPC.

Yet the narrowband LPC still makes occasional pitch and voicing errors. For example a male voice may briefly sound like a female voice due to pitch-doubling. Breath noise is often reproduced as

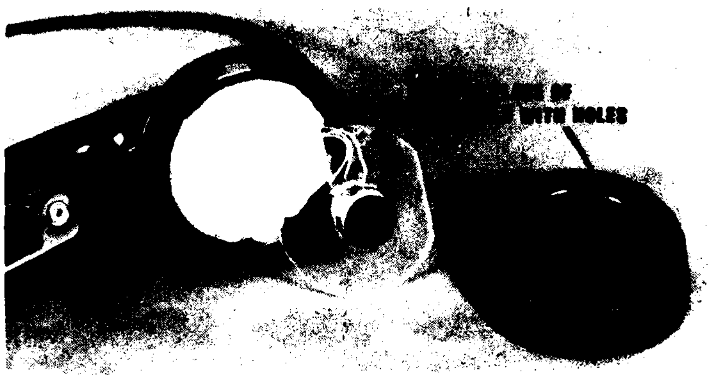


Fig. 23 — Microphone assembly without the standard telephone shield

a snore due to a voicing error. In general these errors are not reflected in the intelligibility scores, but they are reasons the narrowband LPC has not been universally acceptable to general users. Even experienced communicators who accept distorted CB sounds have reservations about the narrowband LPC sounds.

In the past the pitch and voicing algorithms have been custom-made for the capabilities of the particular processor at hand. The algorithms had to be simplified or approximated to make real-time processing possible. Recently a number of powerful processor chips have appeared, and there are still more to come. With the appearance of these chips computationally expensive pitch and voicing algorithms can be incorporated in the narrowband LPC while still keeping it small enough to be a part of the voice terminal.

Currently raw pitch information is often obtained from the average-magnitude difference function performed on low-pass filtered speech via a two-tap inverse filter. The low cutoff frequency is often 800 Hz [2,3]. This cutoff choice is rather irrational, because pitch harmonics of most vowels extend to near 4 kHz and even beyond. The use of a limited bandwidth is one of the more significant reasons for deteriorated pitch estimation when the low-frequency components are absent from the speech (telephone speech) or when low-band speech is contaminated by low-frequency interference.

On the other hand, the voicing decision requires a number of parameters, as it always has, since no single parameter can reliably indicate the state of voicing. For example, the zero-crossing count of the speech waveform is a highly effective means of detecting unvoiced sounds with predominantly high-frequency components, such as /s/, /sh/, and /ch/ but it is virtually useless for detecting unvoiced sounds with predominantly low-frequency components, such as /p/ and breath noise.

We advocate the use of a more powerful analysis not only to derive raw pitch information but also to generate a voicing parameter as a byproduct. This voicing parameter will be used as an adjunct to other voicing parameters in use, not as a substitute for them. Again, a reliable voicing decision needs more than one parameter. The analysis approach we advocate is as follows:

- Raw pitch is estimated from a long-term prediction (the autoregressive analysis) performed on the prediction residual generated by a short-term prediction that produces the LPC coefficients.

- Since the pitch period is unknown a priori, a set of long-term predictors are used. The raw-pitch period is the delay of a long-term predictor that gives the maximum autoregressive coefficient.

- The maximum autoregressive coefficient is a voicing parameter.

The use of the prediction residual for pitch and voicing estimation is logical, because the prediction residual is the ideal excitation signal for the LPC synthesizer. In other words, if the entire prediction residual is fed into the synthesizer, the output equals the input speech that is being analyzed. In essence the excitation signal employed by the narrowband LPC is a drastically simplified model of the prediction residual. Specifically it is either a broadband repetitive signal at the rate of the fundamental pitch frequency—often called the pitch-excitation signal—or random noise. Thus the task of the voicing estimation is to choose which of the two types of waveform best fits the description of the prediction residual.

The most useful descriptors of the prediction residual may be obtained through a first-order regression analysis, which assumes a linear dependency between the prediction residual and the time-delayed prediction residual. Thus

$$\epsilon_t = \theta\epsilon_{t-\tau} + \delta_t, \tag{23}$$

where θ is a constant, ϵ_t is the prediction residual, $\epsilon_{t-\tau}$ is the time-delayed prediction residual, and δ_t is an error whose mean-square value is

$$D(\theta,\tau) = E[(\epsilon_t - \theta\epsilon_{t-\tau})^2], \tag{24}$$

where $E(\ldots)$ implies a short-term averaging. From the theory of least squares, the best estimate of θ is

$$E\{\tilde{\theta}\} = \frac{E[\epsilon_t\epsilon_{t-\tau}]}{E[\epsilon_t^2]}. \tag{25}$$

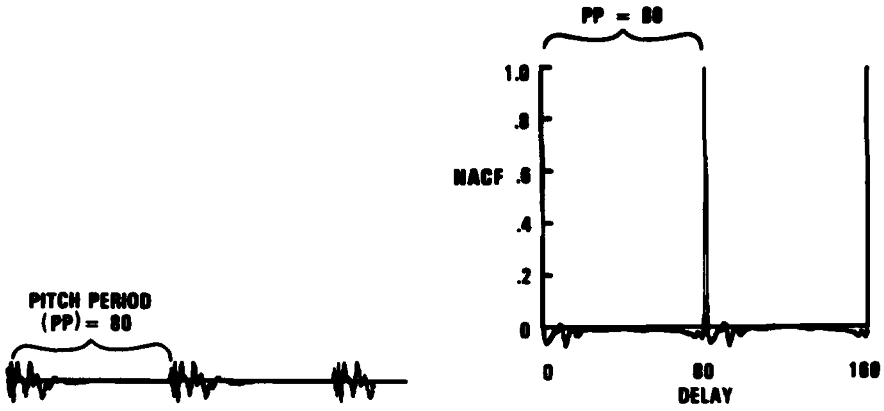
Under the assumption that the residual is of stationary, Eq. (25) equals

$$E\{\tilde{\theta}\} = \frac{E[\epsilon_t\epsilon_{t-\tau}]}{0.5\{E[\epsilon_t^2] + E[\epsilon_{t-\tau}^2]\}}. \tag{26}$$

The delay τ covers the range of pitch periods specified for the narrowband LPC. Equation (26) is a normalized autocorrelation function (NACF). Normalization makes the descriptors of the prediction residual independent of the loudness of the speech. The use of Eq. (26) without normalization has been previously proposed for pitch estimation with nonlinearly transformed input speech [25], because the spectrum of nonlinearly transformed speech is somewhat flat and similar to that of the prediction residual.

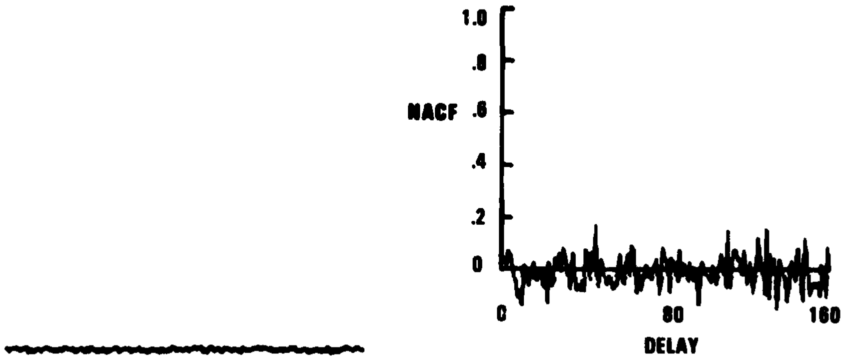
Figure 24 shows the NACF of the pitch-excitation signal used by the narrowband LPC during voiced frames. For this idealized signal the maximum autoregressive coefficient (at $\tau \neq 0$) is 1 0, indicating that the excitation signal is perfectly repetitive. The delay that corresponds to the maximum autoregressive coefficient is the pitch period. Figure 25 is the NACF of the random excitation signal used by the narrow LPC during unvoiced frames. The large peaks are missing, indicating that the excitation signal lacks repetitiveness.

Because the prediction residual of actual speech is far more complex than the narrowband-LPC excitation signal, its NACF is often not as well-defined as that shown in either Fig. 24 or 25. Occasionally the autoregressive coefficient at twice the pitch period may be slightly larger than that at the pitch period. Thus a check for such a case is mandatory. In addition some 'apering of the upper-band residual produces a somewhat cleaner NACF. We used a frequency attenuation of −3 dB at 2 kHz and a gradual attenuation thereafter. Such an attenuation is simply realized by summing two adjacent residual samples (by passing the residual through a two-tap filter with filter weights 1 and 1).



(a) Pitch-excitation signal (b) Autocorrelation function

Fig. 24 — Narrowband-LPC pitch-excitation signal for voiced frames and its autocorrelation function

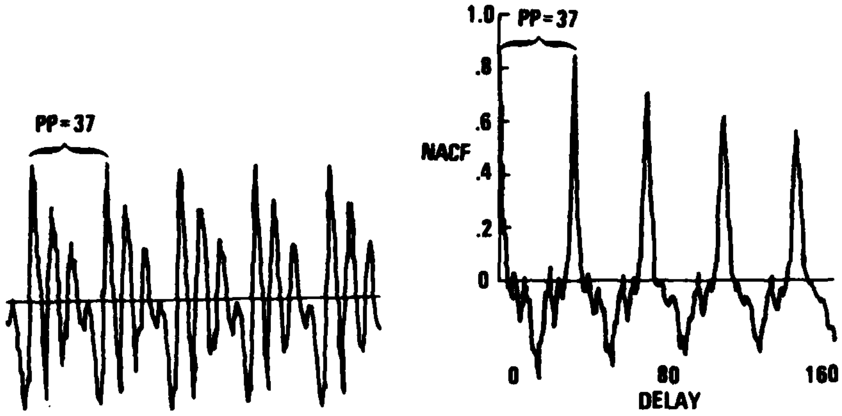


(a) Noise-excitation signal (b) Autocorrelation function

Fig. 25 — Narrowband-LPC noise-excitation signal for unvoiced frames and its autocorrelation function

Pitch and voicing estimation is a decision process whereby the actual residual NACF is classified as either that of a pitch-excitation signal (Fig. 24) or that of a random-excitation signal (Fig. 25). Since the prediction residual is essentially free of formant interference, its NACF is not overly ambiguous, even with noisy speech. Figures 26 through 31 illustrate residual NACFs obtained from various speech segments, including those which often produce erroneous voicing decisions.

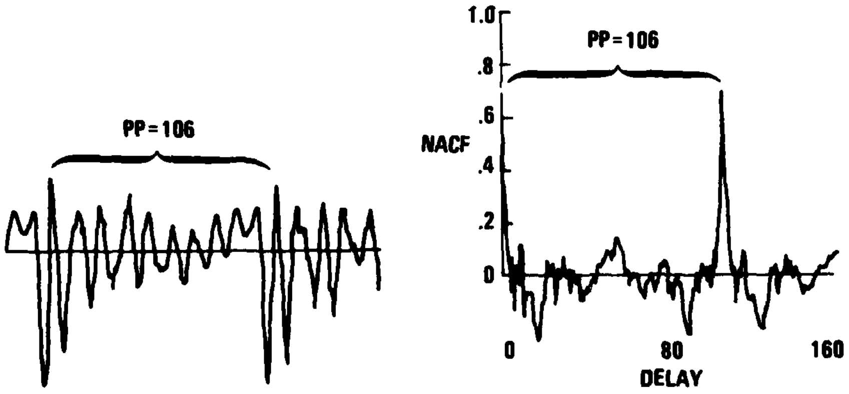
As shown in these examples, the residual NACF provides a quite reliable pitch value and voicing parameter, even for troublesome waveforms. It can be readily adapted for use with the previously mentioned DYPTRACK pitch tracker. In summary, a long-term prediction of the prediction residual is an effective means of pitch and voicing estimation.



(a) Speech waveform

(b) Residual NACF

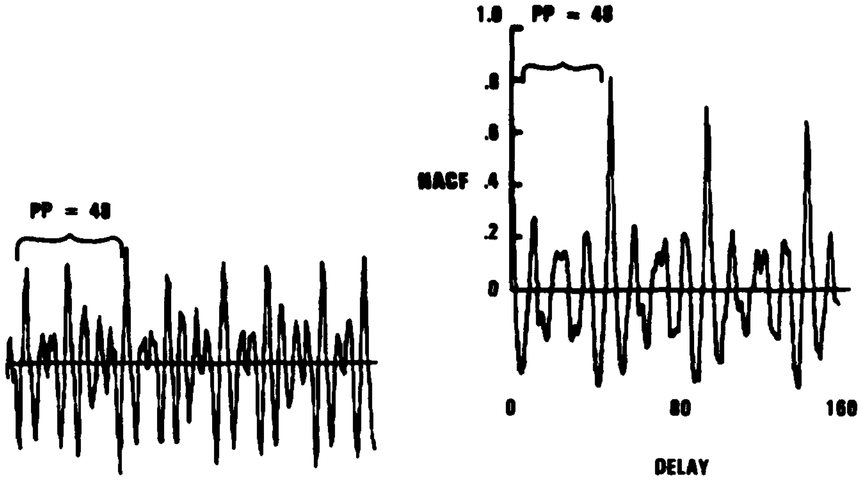
Fig. 26 — Residual NACF obtained from a waveform of high-quality female speech. The residual NACF is clearly that of voiced speech. The delay that corresponds to the first major peak in the residual NACF is the pitch period. Clean speech of this kind does not pose problems for any pitch and voicing estimator.



(a) Speech-plus-noise waveform

(b) Residual NACF

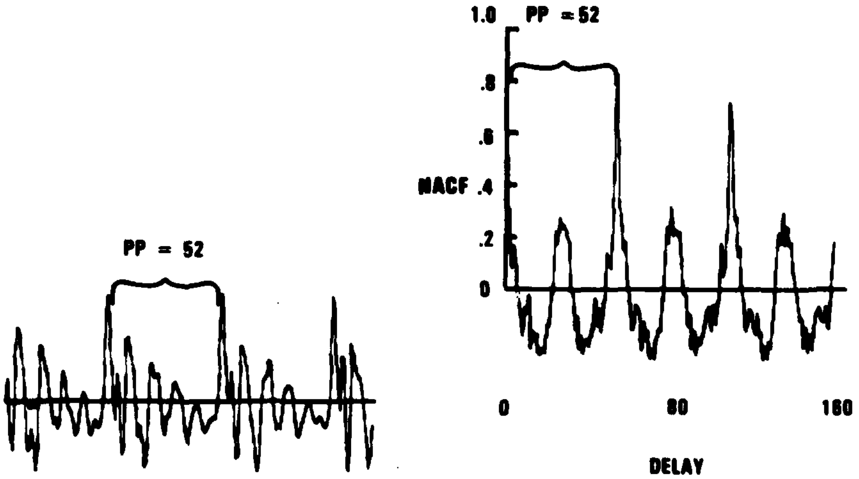
Fig. 27 — Residual NACF obtained from a waveform of high-quality male speech. The comments in Fig. 26 again apply.



(a) Speech-plus-noise waveform

(b) Residual NACF

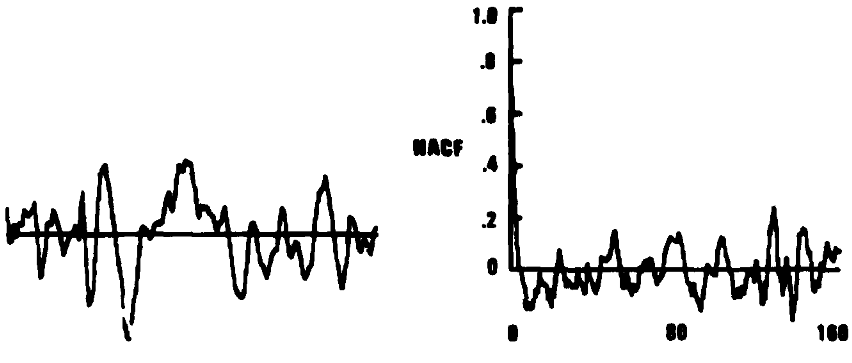
Fig. 28 — Residual NACF obtained from a waveform of male speech corrupted with helicopter noise. The speech waveform has two major peaks in each pitch period because the first-formant frequency is approximately twice the fundamental pitch frequency. Often this type of speech waveform gives rise to pitch doubling. In addition, noise interference can degrade pitch and voicing estimation. The residual NACF resolves this interference, providing an output similar to that with high-quality speech.



(a) Speech-plus-noise waveform

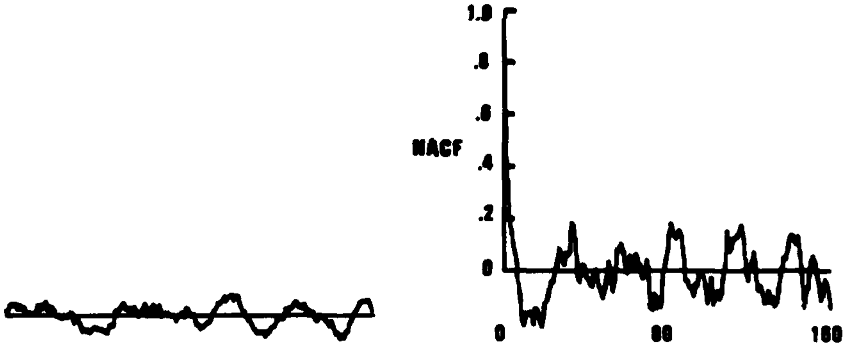
(b) Residual NACF

Fig. 29 — Residual NACF obtained from a waveform of male speech corrupted with M-80 tank noise. The sound pressure level is 112 dB, and the noise has many low-frequency spectral peaks. Yet the residual NACF is expected to give a correct voicing decision and raw-pitch value.



(a) Speech waveform (b) Residual NACF

Fig. 30 — Residual NACF obtained from the unvoiced speech waveform of /p/. Since the spectrum contains predominantly low frequencies, the current voicing decision would often go to "voiced." The residual NACF, however, does not indicate that the sound is voiced, sharp peaks being absent.



(a) Breath-noise waveform (b) Residual NACF

Fig. 31 — Residual NACF obtained from the waveform of breath noise. This is a troublesome waveform which causes a voicing error in the narrowband LPC. The residual NACF, however, as in Fig. 30, is typical of unvoiced sounds.

CONCLUSIONS

The objective of our effort is to improve the narrowband-LPC speech without compromising the existing DoD interoperability requirements on the speech sampling rate, the frame rate, and parameter coding formats. These requirements are expected to remain unaltered for many years. Thus, it is essential to work within these constraints, so that any useful results from these research efforts will benefit the Navy and DoD in general.

The weakness of the narrowband LPC stems from two major sources: it partially eliminates some of the critical features of the original speech in the process of removing speech redundancies, and it notoriously lacks robustness (which lack is an attribute of the least-squares method). Therefore the improvement of narrowband-LPC speech is directly related to the appropriate use of speech preprocessing, proper selection of speech samples for analysis, and suppression of various forms of interference. Covering all of these areas, we have implemented five technical approaches and have tested the results. Each of our modifications substantially improves the performance of the narrowband LPC.

However, because the causes of speech degradation are interrelated and interactive, these improvements must be incorporated in the LPC simultaneously in order to achieve the fullest impact. Furthermore LPC synthesis improvements—which we will present in a forthcoming report—must also be included. Currently we are updating the real-time narrowband LPC to incorporate all of these improvements, and extensive conversational testing will follow.

After nearly a decade of research and development, the narrowband LPC has become a practical means for digitizing speech at low bit rates. Our research effort and similar efforts by other researchers will make the narrowband LPC even more acceptable to general users.

ACKNOWLEGMENTS

Our effort could not have been completed without the help of Ruth Phillips of NRL, who maintained our computer facilities in good working order. Furthermore we appreciate the help of Howard Murphy and Alexander Thornhill, also of NRL, who designed and fabricated the front-end processor for experimentation.

We also thank Robert Martin of the Naval Electronics Systems Command, who supported other activities related to this effort. In particular, he supported Mark Lidd of Signal Processing Solutions, Inc., who generated the software for real-time implementation to test some of our ideas. Finally we gratefully acknowledge the support of the NRL Research Advisory Committee.

REFERENCES

1. A. Schmidt-Nielsen and S.S. Everett, "A Conversational Test for Comparing Voice Systems Using Working Two-Way Communication Links," NRL Report 8583, June 1982.

2. T.E. Tremain, "The Government Standard Linear Predictive Coding Algorithm: LPC-10," Speech Technology 1 (No. 2), 40-49 (Apr. 1982).

3. G.S. Kang, L.J. Fransen, and E.L. Kline, "Mulitrate Processor (MRP) for Digital Voice Communication," NRL Report 8295, Mar. 1979.

4. W.D. Voiers, "Diagnostic Evaluation of Speech Intelligibility," in *Speech Intelligibility and Recognition*, M.E. Hawley, editor, Dowden, Hutchinson, and Kos, Stroudsburg, Pa., 1977.

5. G.W. Hughes and M. Halle, "Spectral properties of fricative consonants," J. Acoust. Soc. Am. **28**, 303-310 (1956).

6. P. Strevens, "Spectra of fricative noise in human speech," Lang. and Speech **3**, 32-49 (1960).

7. V.A. Mann and B. Repp, "Influence of the vocalic context on perception of the [ʃ]-[s] distinction," Percep. and Psychophys. **28**, 213-228 (1980).

8. D.H. Whalen, "Effects of vocalic formant transitions and vowel quality on the English [s]-[š] boundary," J. Acoust. Soc. Am. **69**, 275-282 (1981).

9. M. Halle, G.W. Hughes, and J.-P.A. Radley, "Acoustic properties of stop consonants," J. Acoust. Soc. Am. **29**, 107-116 (1957).

10. K.N. Stevens and S.E. Blumstein, "Invariant cues for place of articulation in stop consonants," J. Acoust. Soc. Am. **64**, 1358-1368 (1978).

11. S.E. Blumstein and K.N. Stevens, "Perceptual invariance and onset spectra for stop consonants in different vowel environments," J. Acoust. Soc. Am. **67**, 648-662 (1980).

12. G.S. Kang, "Application of Linear Prediction Encoding to a Narrowband Voice Digitizer," NRL Report 7774, Oct. 1974.

13. L.R. Rabiner, B. Atal, and M. Sambur, "LPC Prediction Error—Analysis of Its Variation with the Position of the Analysis Frame," IEEE Trans. on Acoust., Speech, and Signal Proc. **ASSP-25**, 434-442 (1977).

14. T.O. Lewis and P.L. Odell, *Estimation in Linear Models,* Prentice-Hall, Englewood Cliffs, N.J., 1971.

15. E. Matsui, T. Nakajima, and H. Omura, "An Adaptive Method of Speech Analysis Based on Kalman Filtering Theory," Electrotech. Res. Inst. Rep. **36** (No. 3), 210-219 (1972) (in Japanese).

16. C.J. Atkinson, "A Study of Vocal Responses During Controlled Aural Stimulation," J. Speech and Hear. Dis. **17**, 419-426 (1952).

17. G.M. Siegel and H.L. Pick, Jr., "Auditory feedback in the regulation of voice," J. Acoust. Soc. Am. **56**, 1618-1624 (1974).

18. A. Schmidt-Nielsen and D.C. Coulter, "Effect of Modest Sidetone Delays in Modifying Talker Rates and Articulation in a Communications Task," in *Speech Communication Papers Presented at the 97th Meeting of the Acoustical Society of America,* J.J. Wolf and D. Klatt, editors, June 1979.

19. Ketron, Inc., "ANDVT Microphone and Audio System Study, Final Report," prepared for the Naval Electronics Systems Command, Washington, D.C., 1978.

20. H. Olson, "Gradient Microphones," J. Acoust. Soc. Am. **17**, 192-198 (1946).

21. B. Widrow, J.R. Glover, Jr., J.M. McCool, J. Kaunitz, C.S. Williams, R.H. Hearn, J.R. Zeidler, E. Dong, Jr., and R.C. Goodlin, "Adaptive Noise Cancelling: Principles and Applications," Proc. IEEE **63**, 1692-1716 (1975).

22. S:F. Boll and D.C. Pulsipher, "Suppression of Acoustic Noise in Speech Using Two Microphone Adaptive Noise Cancellation," IEEE Trans. on Acoust., Speech, and Signal Proc. **ASSP-28**, 752-753 (1980).

23. S.F. Boll, "A Spectral Subtraction Algorithm for Suppression of Noise in Speech," pp. 200-203 in *Proceedings of the 1979 IEEE International Conference on Acoustics, Speech, and Signal Processing.*

24. W.A. Blankenship and W.R. Bauer, "DYPTRACK, A Noise-Tolerant Pitch Tracker," internal memorandum of a DoD agency, 1973. (The essential elements of DYPTRACK may be found in Appendix C of Ref. 2.)

25. L.R. Rabiner, "On the Use of Autocorrelation Analysis for Pitch Detection," IEEE Trans. on Acoust., Speech, and Signal Proc. **ASSP-25**, 24-33 (1977).

DAT
ILM